NASA

Technical Memorandum 82119

# CONTROL OF MULTIPLE EXPONENTIAL SERVERS WITH APPLICATION TO COMPUTER SYSTEMS

Ronald LeRoy Larsen

APRIL 1981

National Aeronautics and
Space Administration

**Goddard Space Flight Center**
Greenbelt, Maryland 20771

# CONTROL OF MULTIPLE EXPONENTIAL SERVERS WITH APPLICATION TO COMPUTER SYSTEMS

by
Ronald LeRoy Larsen

Dissertation submitted to the Faculty of the Graduate School
of the University of Maryland in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
1981

## ABSTRACT

A class of dynamic control policies is defined for scheduling customers from a Poisson source on a set of exponential servers with dissimilar service rates. A fastest-to-slowest ordering is imposed on the servers, and they are invoked in response to instantaneous system loading as measured by the length of the queue of waiting customers. Markov chain analysis is employed to analyze the performance of the control policies and to develop optimality criteria. It is shown for the two-server case, and believed to be true in general, that probabilistic control policies are suboptimal to minimize the mean number of customers in the system, and that the optimal policy is in a restricted class of policies referred to as "threshold queueing" policies. In a threshold queueing policy, specific queue lengths are identified as "thresholds" beyond which an additional (fastest idle) server is invoked. A new server remains busy until it completes service on a customer and the queue length is less than its invocation threshold. Optimality conditions are derived, and an approximation to the optimum policy is analyzed. For most operational applications, a very simple approximation of the optimal threshold value suffices.

Extensions to the basic policy involving different objectives are considered, including a treatment of the n-server case, the finite queue situation, and inverse ordering (slowest-to-fastest) of servers. Several applications of threshold queueing policies in computer and communications systems are presented. It is concluded that threshold queueing policies with easily approximated thresholds provide near-optimal control of multiple exponential servers with dissimilar service rates, and that these policies can be readily applied to improve the performance of contemporary computer and communications systems.

Acknowledgments

I am deeply indebted to my advisor, Dr. Ashok K. Agrawala, for the years of intellectual stimulation and personal support he has provided me. His curiosity, vision, pragmatism, and friendship have contributed immeasurably to the completion of this dissertation. I would also like to thank Drs. Satish Tripathi and Lawrence Dowdy, as well as Dr. Arnold Greenland at George Mason University, for the valuable discussions we have had and the time they have invested in critiquing drafts of this dissertation.

My wife, Lida, deserves special recognition for the many personal sacrifices she has made during the preparation of this dissertation. She has been a constant source of support in innumerable, invaluable ways, and her contributions are greatly appreciated.

Many of the results reported in this dissertation were obtained with the assistance of MACSYMA, which was developed by the Mathlab group of the Laboratory for Computer Science at the Massachusetts Institute of Technology with the support of the National Aeronautics and Space Administration under Grant NSG 1323.

TABLE OF CONTENTS

Appendix

LIST OF FIGURES

# CHAPTER 1

# INTRODUCTION

Queueing theory research, in response to the analytic requirements of computer and communications systems, has developed in many directions, expanding to address different arrival and service patterns, new configurations of queues and servers, and alternative means of controlling the internal operation of queueing models. Multiple server configurations have attracted substantial attention due to their wide applicability to computer and communications systems analysis. The literature on this class of problems is vast, addressing various arrival and service distributions, and alternative scheduling disciplines, but generally limited to configurations in which all servers have identical service rates (the M/M/m, G/M/m, and related problems). Although it simplifies the analysis of the multiple server system, this assumption is frequently violated in operational systems. In any system where the servers are humans, for example banks and grocery stores, homogeneous performance characteristics among the tellers or clerks would be considered quite unusual. In computer and communications systems, a sufficient spectrum of capability exists within the classes of most devices (cpu's, storage devices, communication lines, etc.) to make the feasibility of, and likelihood of encountering, heterogeneous combinations of devices within a class quite high. This might occur, for example, in a communications network supporting communication channels of differing transmission rates, in a multiprogramming computer system which spools its output for printing on a set of line printers of differing speeds, or in scheduling jobs or transactions on functionally equivalent processors of a local computer network. The analysis and control of heterogeneous multiple server queueing systems such as these has not been substantially addressed to date, and forms the primary focus of this dissertation.

The basic problem to be considered is illustrated in figure 1-1.

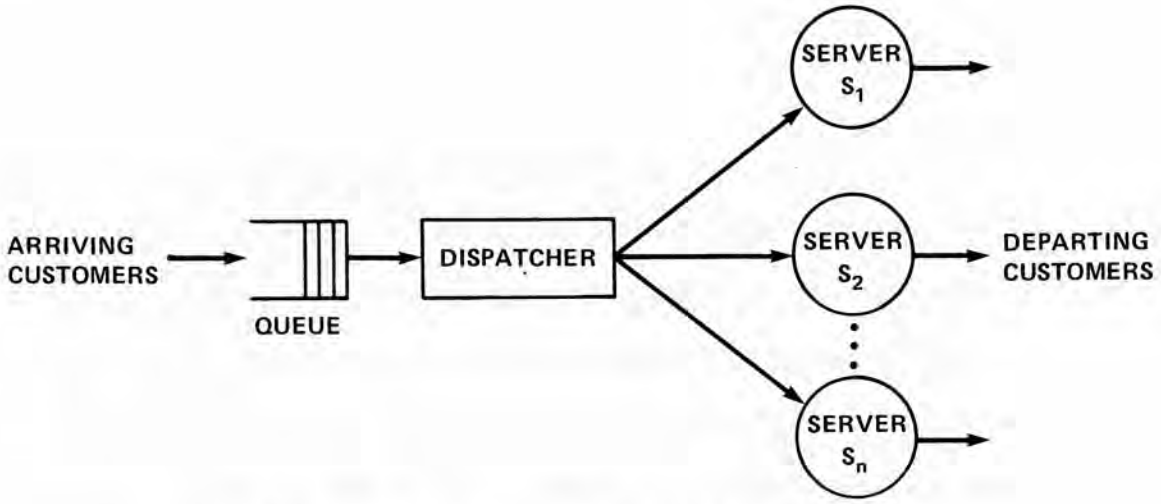



Figure 1-1. Heterogeneous multiple server system

Rather than an assumption of homogeneous exponential servers, it is assumed that the service rates $\mu_i$ of the various servers $S_i$ are different. For the sake of clarity, it is generally assumed that $\mu_1 \geq \mu_2 \geq \ldots \geq \mu_n$. Customers arrive from a Poisson source of rate $\lambda$ and immediately join the queue. A dispatching (control) function controls the access of customers to the various (exponential) servers, and following completion of service, customers depart the system. The dispatcher makes its scheduling decisions on the basis of the busy/idle status of the servers and the length of the queue of waiting customers. In this dissertation, rigorous treatment of the system of figure 1-1 is limited to the case of two exponential servers with dissimilar service rates. A scheduling discipline referred to as Threshold Queueing is defined and analyzed. In a threshold queueing discipline, customers are preferentially routed to the faster server. Customers are allowed to queue up while the slower server remains idle until the queue size reaches a certain "threshold" value, at which point a customer is removed from the queue and sent to the slower server for service. The threshold value thereby becomes a critical control parameter affecting system performance, and facilitating optimal system control. The primary performance parameter is the mean number of customers in the system, $\overline{N}$. Optimization of the two-server heterogeneous server problem is considered over an infinite time horizon with an average

cost criterion ($\bar{N}$). Linear holding and service costs are considered, and it is assumed that there is no additional cost incurred to turn on or turn off a server.

While the queueing theory literature does not address the analysis of the threshold queueing discipline, many related topics are addressed. Chapter 2 presents a brief survey of the relevant queueing theory literature. This chapter begins with a broad and general perspective on the queueing theory literature, citing some of the relevant contemporary texts and survey papers. The focus is narrowed through the course of the chapter, considering how queueing analysis has been applied to the analysis of computer and communication systems, then considering some of the contributions of the operations research and management science literature to the control of queueing systems, and finally focussing in on several papers which are directly related to the threshold queueing problem.

In chapter 3, the general n-server problem is formulated as a Markovian decision process. Control policies for the two-server case are characterized in terms of trajectories within the system state space, and the structure of this state space is used to eliminate clearly sub-optimal classes of policies. As a result, the threshold queueing discipline emerges as being worthy of analysis, and Markov chain analysis is utilized to construct the steady state performance equations. An algorithmic solution to the steady state equations is presented, and the chapter concludes with a closed-form solution to these equations.

Chapter 4 considers optimization of the performance of the two-server threshold queueing discipline, where minimizing $\bar{N}$ is the optimization criterion. A major result of this chapter is a proof that the optimal threshold queueing control policy is deterministic rather than probabilistic. This result significantly simplifies both the operation of the system and the computation of its performance. The parameter space is partitioned into regions within which the optimal control parameter (threshold size) is constant, and a linear approximation to the boundaries of these regions is computed and proposed for the control policy. Chapter 4 concludes with an analysis of the optimized performance of the

threshold queueing discipline and a conjecture that probabilistic control policies in general are sub-optimal for the n-server problem.

In chapter 5, applications of the basic threshold queueing discipline are considered, and several extensions are treated. The basic applications considered include controlling the scheduling of messages over heterogeneous communication lines and controlling the use of line printers in a multiprogramming system. Two extensions are considered: slow server preference, and finite queue capacity. In some systems, either by design or constraint, the primary, or preferred, server is slower than the secondary one. This situation is captured by the basic threshold queueing solution, but the optimization results no longer apply, and the objective is typically to strive toward consistent performance rather than optimal performance. A time-sharing example is used to illustrate this case. The existence of a finite rather than infinite queue requires a minor reformulation of the steady state performance equations. Following this derivation, an onboard data processing application for an earth-resources type of spacecraft is considered as an application for the finite queue case. Chapter 5 concludes with a discussion of the n-server case and an outline of an algorithmic solution for the steady state performance.

Chapter 6 summarizes the major contributions of this dissertation and identifies potentially fruitful topics for future research.

# CHAPTER 2

# LITERATURE SURVEY

Advances in queueing theory occur in a diverse set of disciplines, representing the breadth of application to which queueing theory has been a valuable analysis tool. The literature is quite rich. In this chapter, relevant contributions to the topic of this dissertation are surveyed, putting the problem addressed herein in its proper perspective with respect to other work of a similar nature.

Recent theoretical advances within operations research have addressed adaptively controlling queueing models to maximize an objective function, and have frequently exploited the notion of Markov decision processes ([HOWAR71] and [DERMC70]) in so doing. This approach has yet to find wide application in computer science. It is utilized in the analysis presented in this dissertation, and has the potential of broad applicability to computer science problems, particularly in developing dynamic system control policies.

## 2.1 Analysis of Queueing Systems

Queueing theory has a long history of development, from Erlang's work early in this century to the present, resulting in a vast literature. The theory and recent applications of queues have been surveyed by Rosenshine [ROSEM75]. His article utilizes the following six descriptors of a queueing system:

1. Distribution of arrivals
2. Service time distribution
3. Number of service channels
4. Queueing discipline
5. System capacity
6. Number of service stages

and presents a detailed list of queueing models for which equilibrium solutions are known. In a series of articles and a book ([ALLEA75], [ALLEA78], [ALLEA80]), Allen treats the fundamentals of queueing theory, particularly as it applies to computer systems analysis. His more recent paper, [ALLEA80], establishes a perspective on queueing theory, showing its relationship to other analysis techniques, such as rules of thumb, linear projection, simulation, and benchmarking. This latter paper goes on to survey major queueing theory results which relate specifically to computer systems modeling, such as open queueing models (M/M/c), finite population models of interactive computer systems, and the central server model. A more detailed introduction to the application of queueing analysis in computer systems is provided by Coffman and Denning in chapter 4 of [COFFE73], and by numerous authors in the special September 1978 issue of ACM Computing Surveys. Kleinrock's two volume set, [KLEIL75] and [KLEIL76], provides one of the most complete treatments of the theory and application of queueing theory to computer and communications system analysis. In an earlier work, [KLEIL70], Kleinrock surveyed analytical methods particularly suited to the analysis and optimization of the performance of computer communication networks of a given topology, under constraints, and with respect to a given objective function. Of prime concern in this paper is the assignment of channel capacities to communication links of an Arpanet-like network. Subsequently, Kobayashi and Konheim in [KOBAH77] reviewed queueing analysis techniques essential to the analysis of computer communication systems, focussing on buffer storage problems, multiplexing techniques, and network configurations. This paper also includes an extensive bibliography. Typical of the types of analyses utilizing queueing theory for the analysis of computer communications systems is [AGNEC76], a study of adaptive routing algorithms. In this paper, necessary conditions for optimal routing are derived using a Lagrangian, and it is shown that the equilibrium property of adaptive routing (the average travel time from a given node to a given destination is equal for all routes which have traffic assigned to them and less than the travel time for all routes

with no traffic) is inconsistent with the necessary conditions for optimal routing. The problem is that the intuitive algorithms for adaptive routing disregard potential future effects on the network of present actions. The inconsistency is removed by changing a linear bias term to a quadratic one, yielding a 5-10% performance improvement. The analytical results are verified through simulation.

Queueing theory has been widely used to analyze the performance of multiprogramming systems. Gaver [GAVED67] considers a single cpu model with several I/O devices, all identical in speed, and a fixed degree of multiprogramming. Jobs in the system are considered to be in one of four states: being processed by the cpu, waiting for the cpu, engaging in I/O, or waiting for I/O. Queueing for I/O occurs only if all I/O devices are busy and a job requests I/O. Gaver computes the cpu utilization for a system in which all I/O devices are characterized by exponential service distributions, and the cpu by constant, exponential, Gamma, hyperexponential, and hypoexponential service distributions. While analytically interesting, this model suffers from lack of reasonable correlation to the operation of most third generation multiprogramming systems. A more accurate model was provided by the central server model, shown in figure 2.1-1, which provided for probabilistic routing and queueing of requests for each I/O device independently.

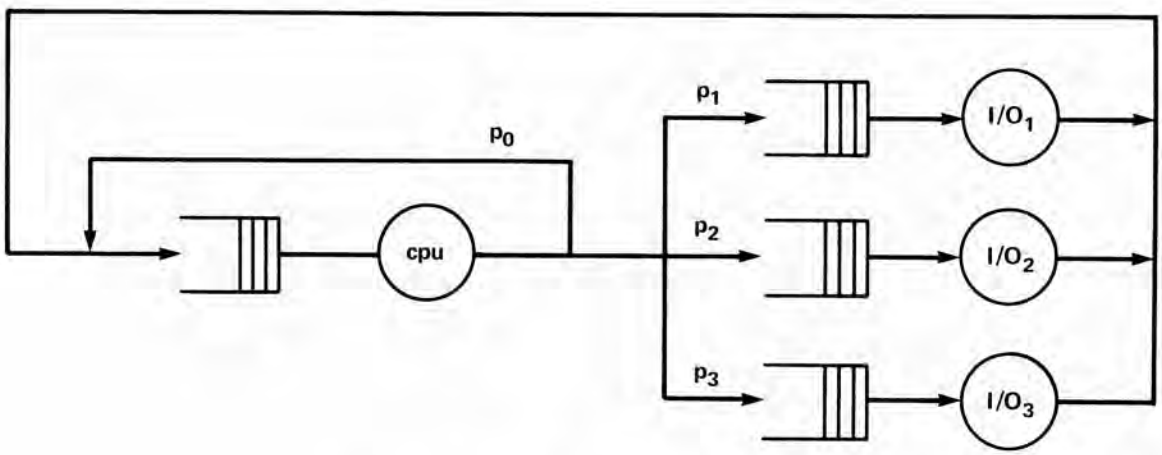



Figure 2.1-1. Central server model

Buzen's development of computational algorithms for the central server model [BUZEJ73] did much to further the use of queueing theory for the analysis of multiprogramming systems. Other examples include [STRAJ74], [BOYSJ75], and [CHENP75]. Strauss, in [STRAJ74] developed a "cyclic-server multiprogramming system model" to model the performance of an IBM S/360 OS/MFT system running the HASP Execution Task Monitor. As in [AGNEC76], model validation is performed through simulation. A similar model is developed in [BOYSJ75] for an interactive system with demand paging under a fixed degree of multiprogramming and a heavy load. In this paper, exponential and deterministic service times are considered and equations are derived for cpu utilization, average response time, and throughput. Chen, in [CHENP75], extends Buzen's work to consider state dependent routing probabilities for an interactive system with swapping. He presents an iterative algorithm utilizing Buzen's methods and shows convergence occurs regardless of the starting point. In a recent paper by Towsley, [TOWSD80], state-dependent routing is treated in a more general context for a closed queueing network. The routing functions are allowed to be a rational function of the queue lengths of downstream queues, and it is shown that if a network without state-dependent routing has a product form solution, then that network with state-dependent routing will also have a product form solution. As an example, Towsley considers state-dependent routing to two identical, parallel processors.

Multiprocessor systems have similarly been modelled using queueing theory. Sastry and Kain [SASTK75] present analytic results for a model of p equivalent processors which all have access to m memory modules, all simultaneously accessible with equal access times. Instruction execution rate is used as the performance measure. Simulation results are presented to validate the model as applied to a Univac 1100-series machine. Chandy, Sauer, and Browne [CHANK75] briefly overview analytical techniques appropriate for parallel computer systems, including precedence graphs and queueing theory.

The success of queueing theory in general, and network models in particular, to the analysis of the performance of computer and communications systems has led to the development of interactive systems which enable users to construct and solve queueing models. Chandy, Keller and Browne describe one such system in [CHANK72].

## 2.2 Control of Queueing Systems

Section 2.1 very briefly surveyed the application of queueing theory to the analysis of computer and communications systems. With rare exception, most of these applications address the analysis of a static system configuration operating under a non-adaptive control discipline. The purpose of most of these analyses is to model the steady state performance of an existing system, calibrate the model to data taken from the subject system, and then perturb the system configuration to accommodate a desired change – a faster cpu, an additional disk, etc. The perturbed model is then used to predict the performance of the actual system under the configuration change. The additional power of queueing models to explore dynamic (adaptive) control policies has been recognized by some researchers. In this section, a brief review of this literature is presented, as well as a few examples which illustrate the application of this approach to computer systems analysis.

Stidham and Prabhu [STIDS73] argue for the development of a unified theory of queueing control, and suggest research emphasizing a common structure, utilizing classical control theory, particularly stochastic control. They present a good survey and unification of work on queueing control problems, categorizing the work by: (1) system structure (e.g., M/M/1), (2) decision variables, (3) admissible decision epochs, (4) costs, (5) the objective function, and (6) the time horizon. Crabill, Gross, and Magazine [CRABT77] review the literature on optimal design and control of queues and present a taxonomy and classified bibliography. The essence of their taxonomy is shown below:

1. Static (Design) Models
2. Dynamic (Control) Models
    - 2.1 Arrival Process Control
        - 2.1.1 Accept or Reject Customer
        - 2.1.2 Adjust Mean Arrival Rate
        - 2.1.3 Customer-exercised Control
        - 2.1.4 Self vs. Social Optimization
        - 2.1.5 Rejection Times
    - 2.2 Service Process Control
        - 2.2.1 Varying Number of Servers
            - 2.2.1.1 Single-server On-off
            - 2.2.1.2 More than One Server
            - 2.2.1.3 Dispatching Times
        - 2.2.2 Varying the Service Rate
3. Control of Queue Discipline
    - 3.1 Priority Models
    - 3.2 Scheduling Models
    - 3.3 Allocation of Customers to Multiple Servers

Most of the applications of queueing theory to computer systems analysis have been in the category of static (design) models. Leroudier and Potier [LEROJ76], however, present an intuitively based adaptive controller which adjusts the degree of multiprogramming in the central server model. Utilizing the taxonomy of [CRABT77], this paper would be classified into section 3.2, Scheduling Models. Simulation results are presented to show the degree of improvement available using the adaptive controller. Optimization is performed based on the following two "principles":

(1) In a multiprogrammed virtual memory computer, optimum performance is achieved if the utilization of the secondary memory remains in the 50% region, and

(2) At the optimum working point the sensitivity of the degree of multiprogramming to the utilization of the secondary memory is minimum.

Denning and Kahn [DENNP76] perform an analysis via simulation of a similar intuitive algorithm. Their analysis explores the balanced relationship between interpagefault lifetime (L) and page swap time (S), and, in particular, shows that if L is only slightly greater then S, then near optimum performance results. This algorithm outperforms principle 1 listed above. While both of these papers approach the notion of utilizing queueing models to develop adaptive control policies, both rely on intuitive arguments and do not rigorously formulate the decision process.

Reiser and Konheim [REISM76] analyze a two server cyclic queueing system preceded by an infinite capacity queue, or buffer. Only a limited number of customers are allowed in the two queues, subsequent arrivals being held in the buffer until a departure frees a position in the cyclic arrangement of queues. Solution requires finding the roots of a characteristic equation, for which a solution algorithm is provided. Given a set of operating parameters (the degree of multiprogramming, service rates, etc.) an algorithmic solution for the model can be computed. While this paper treats the problem as a static design problem, it would be a relatively minor step to generalize it to solve for an optimal degree of multiprogramming. One might then be able to develop an optimal controller for the system to adjust the degree of multiprogramming as a function of system operating parameters.

Trivedi and Wagner [TRIVK80] optimize the operating parameters of the central server model. The treatment is, again, a static design one in which an optimal combination of cpu speed, secondary storage device capacities, and the allocation of files to secondary storage devices is sought. It is shown that there exists a unique global maximum, and a technique is presented to reduce the dimensionality of the problem. System cost is treated as the sum of component costs, where cpu cost is modelled as a power

function of speed, and secondary storage costs are considered to be a linear function of capacity. This is another example of a problem formulated as a static design problem which may be amenable to dynamic control through a decision process formulation.

Multi-server systems have provided a focus for much of the research on the optimal control of queueing systems. Magazine [MAGAM71], for example, proves the existence of an optimal policy for the control of multi-server systems under periodic review. His treatment considers multiple identical exponential servers, Poisson arrivals, finite system capacity, and decision points spaced in equal time intervals. A constant cost is incurred to open and close a server; server operation costs and customer holding costs are both assumed to be linear. Bell [BELLC75] goes on to investigate the form of the optimal policy when the number of servers working can be adjusted at arrival or service completion epochs. He shows that with an average cost criterion over an infinite time horizon, an optimal policy may shut down a server even when customers are waiting for service.

Weber [WEBER80] shows by a dynamic programming type of argument that the mean queueing time for a G/GI/m system is a nonincreasing and convex function of the number of servers, m. This means that the improvement yielded by adding two additional servers will be less than twice the improvement realized by adding one additional server. The significance of this particular result is that it implies the optimality of marginal analysis in allocating multiple servers among several service facilities where the objective is to minimize the sum of the mean queueing times at the facilities.

Levy and Uri [LEVYY76] analyze the performance of an M/M/s queueing system in which a server which completes a service period to find no customers waiting becomes idle for an exponentially distributed period of time called a "vacation". Two alternatives are considered for server behavior at the end of a vacation. If no customers are waiting, it may become idle once again for an exponentially distributed period of time. Alternatively, the server will wait for the next customer arrival and immediately begin a service

period. Formulas for the distribution of the number of busy servers and for the mean number of customers in the system are presented. It is shown numerically that the mean number of customers in the system is very nearly a linear function of the mean length of a vacation.

The complexity of precise analysis of multiple server systems has resulted in many cases in the development of simpler approximations to, bounds on, and inequalities between system performance parameters. Sakasegawa [SAKAH77] notes that a formula for the mean queue length of the GI/G/1 queue is "not near at hand," and derives an approximate formula for the mean wait time of an M/G/1 queue and a general approximation for the mean queue length which is shown numerically to be not too bad for M/M/s, $E_2$/M/s, $E_4$/M/s, $E_{10}$/M/s, M/D/s, D/M/s, and $E_2/E_2$/s queues. Boxma, Cohen, and Huffels [BOXMO79] develop an approximate formula for the mean waiting time in an M/G/s queue, with particular attention given to the G = D case, which is non-exhaustively shown to be generally accurate to within 2 - 3%. Sobel [SOBEM80] and Heyman [HEYMD80] consider the G/G/c/N system and derive lower and upper bounds on the long term probability that the facility is full. It is also shown by numerical example in these papers that the lower bound is, in fact, a good approximation for a system with at least two servers under heavy loading.

An alternative to considering multiple server systems is to consider an M/G/1 (or, more generally, G/G/1) system in which the service rate can be switched among several values as a function of system load. A good introduction to M/G/1 and GI/M/1 queueing systems can be found in [BHATU68]. Yadin and Naor [YADIM67] formulate the control problem for an M/G/1 system in which the service rates can be varied as a function of the queue length and the recent history of the system. The cost function includes linear holding and service costs, and fixed costs for changing the service rate. As their treatment of the control problem is quite general and similar to the system analyzed in this

dissertation, it will be sketched here. Alternative service rates are specified by a vector $\vec{\mu} = \{\mu_0, \mu_1, \ldots, \mu_k, \ldots\}$ in which $\mu_0 = 0$ and $\mu_{k+1} > \mu_k$. Two control vectors $\vec{R} = \{R_1, R_2, \ldots\}$ and $\vec{S} = \{S_0, S_1, \ldots\}$ are defined, for which $R_{k+1} > R_k$, $S_{k+1} > S_k$, $R_{k+1} > S_k$, and $S_0 = 0$. The $\vec{R}$ vector specifies queue sizes at which the service rate should be incremented, and the $\vec{S}$ vector specifies queue sizes at which the service rate should be decremented. When there are less than $R_1$ customers in the system, for example, the service rate will be $\mu_0$. When the queue size reaches $R_1$, the service rate is incremented to $\mu_1$; when it falls to $S_0$, the service rate is decremented to $\mu_0$. The state of the system is given by (k, i), in which k represents the system "phase", meaning the service rate is $\mu_k$, and the queue size is i. States (k, $R_k$) and (k, $S_k$) represent entry points into phase k; (k, $R_{k+1} - 1$) and (k, $S_{k-1} + 1$) are exit points. Non-zero service rate switching costs lead to "hysteresis policies," in which $R_k > S_k$, capturing the recent history of operation of the queue. Some basic expressions are developed for the expected queue size and related performance parameters. While optimization is briefly considered, combinatoric complexity is cited as making it nearly impossible unless the specific problem can be constrained. Shanthikumar [SHANJ79] discusses such a constrained problem in which a customer's service time is dependent on whether any newly arrived customers are behind it at the onset of service. Server utilization and the Laplace transform of the distribution of customer wait time are derived for M/(G, G)/1, M/(M, G)/1, and M/(M, M)/1 systems. Cohen [COHEJ76] evaluates an M/G/1 system with two service rates, investigating optimal criteria for switching the service rate as a function of the virtual waiting time. The cost function includes the cost of maintaining the service at a given level, switching costs, and linear customer holding costs. Bell [BELLC80] utilizes the basic formulation developed in [YADIM67] to consider the optimal operation of an M/M/2 system with removable servers. He shows that the optimal form of policy for an average cost criterion over an infinite time horizon is a hysteresis policy, characterized by $\vec{R} = \{R_1, R_2\}$ and $\vec{S} = \{S_0, S_1\}$, where:

(1) $R_i \geq i$, $i = 1, 2$

(2) $R_2 \geq R_1$

(3) $-1 \leq S_0 \leq S_1$

(4) $R_2 > S_1$

(5) $S_0 \leq 0$.

A negative value for $S_i$ means that the number of working servers is never adjusted downward to i. It is also shown in this paper that there exists a queue size above which both servers should be left on. This group of papers ([YADIM67], [SHANJ79], [COHEJ76], and [BELLC80]) relates most closely to the work presented in this dissertation.

A closely related body of research addresses the control of the flow of customers to multiple servers, each with its own queue. This is the load balancing problem. Chen [CHENP73] develops an algorithm for the allocation of files among storage devices which can also be used to probabilistically direct customers in an optimal fashion to M/M/1 servers in a multiple server system. Figure 2.2-1 illustrates this problem. Chen's algorithm determines the optimal dispatching probabilities $p_i$.

Fayolle and Robin [FAYOG76] consider two coupled exponential servers with Poisson arrivals and finite queues in which jockeying, or the movement of customers from one queue to the other, is allowed. The cost structure considered includes linear holding

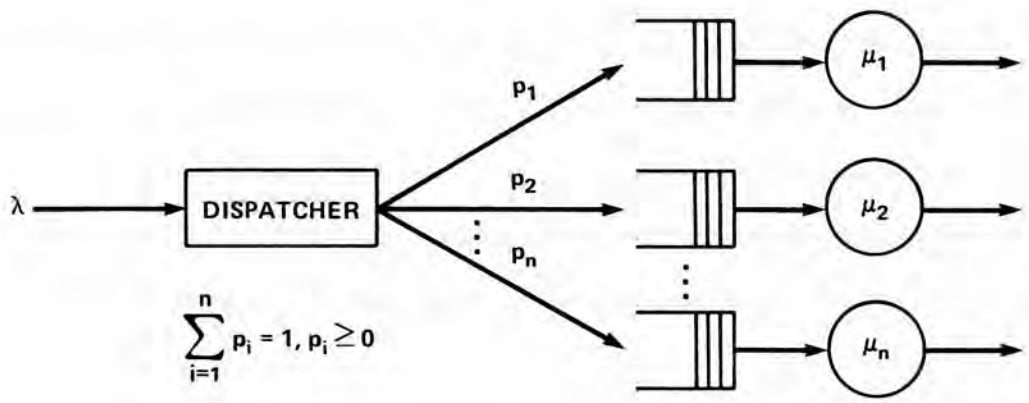



Figure 2.2-1. Probabilistic dispatching in a multi-server system

costs and fixed customer transfer costs. The existence and uniqueness of an optimal policy is proven and characterized by a partition of the state space into three regions, representing the three viable control options:

(1) do not transfer

(2) transfer a customer from queue 1 to queue 2

(3) transfer a customer from queue 2 to queue 1.

A numerical example is presented for the infinite horizon, discounted cost case, but no performance measures are computed, so the amount of improvement provided by this policy is unknown.

Winston [WINSW77] considers a discrete time queueing system with multiple servers of varying rates and multiple customer classes. At the beginning of each period, customers may be reassigned among the servers in a preemptive fashion. Winston's treatment of the problem considers an infinite time horizon, yielding an uncountable state space. It is shown that an optimal policy will always assign customers from a class with longer service time requirements to faster servers.

Chow and Kohler [CHOWY79] consider non-deterministic and deterministic load balancing policies for a heterogeneous multiple server system with the objective of reducing the mean customer time in the system. The probabilistic policy considered is the proportional branching policy which routes customers to servers with probabilities in direct proportion to the service rates of the respective servers. This policy is inferior to the probabilistic policy developed in [CHENP73]. Three deterministic load-balancing policies are considered:

(1) Minimum expected turnaround time for new arrival

(2) Minimum expected time to complete service on all jobs in system plus new arrival

(3) Maximum throughput during next interarrival period.

It is shown numerically that the third deterministic policy minimizes the mean customer time in the system among the policies considered, and a conjecture is made that this policy is optimal. An approximate recursive solution is employed which is not easily generalizable to more than two servers.

Ricart [RICAG80] presents a simulation-based analysis of decision policies for heterogeneous multiple server systems. Two classes of problems are considered: (1) one arrival process with multiple servers, and (2) multiple independent arrival processes with multiple servers. With one arrival process, routing decisions are made as a function of arrival and service rates, server backlogs, past routing decisions, and any knowledge which is available of specific customer service requirements. Constant and exponential customer service time requirements are considered, and five routing algorithms are evaluated, ranging from proportional probabilistic routing, through adaptive routing algorithms which anticipate future loading conditions. The performance of these algorithms is compared to the optimal system performance, showing that the adaptive routing algorithms result in very close to optimal performance. The effect of delayed information on adaptive routing algorithms is also considered. When multiple arrival processes are considered, the routing problem has an additional dimension of complexity. Independent controllers must make routing decisions with potentially limited information regarding the decisions of other controllers. This introduces the notion of an information structure governing the sharing of information among controllers. Ricart compares twenty-seven different policies for the control of these kinds of systems and reports a wide spectrum of performance results.

With respect to the taxonomy developed in [CRABT77], the main focus of this dissertation is on Dynamic (Control) Models, and, more specifically, on Service Process Control rather than Arrival Process Control. In this section a sample of the literature on Service Process Control has been surveyed, with particular emphasis on those policies

directly related to the policy to be developed in chapter 3 of this dissertation. In addition, a brief review of the literature on the Control of Queue Discipline, Allocation of Customers to Multiple Servers, has been presented for completeness.

# CHAPTER 3

# NON-PREEMPTIVE SCHEDULING OF MULTIPLE EXPONENTIAL SERVERS WITH DISSIMILAR SERVICE RATES

The queueing system to be considered here consists of a single Poisson arrival source of rate $\lambda$, an unbounded queue, and multiple (n) exponential servers of rates $\mu_1$, $\mu_2$, . . . $\mu_n$, respectively. All customers are considered identical in that they can receive service from any server. Arriving customers join the queue. Available (idle) servers may remove a customer from the queue, serve that customer, and thereby enable the customer to depart the system. The overall system is depicted in figure 3-1.

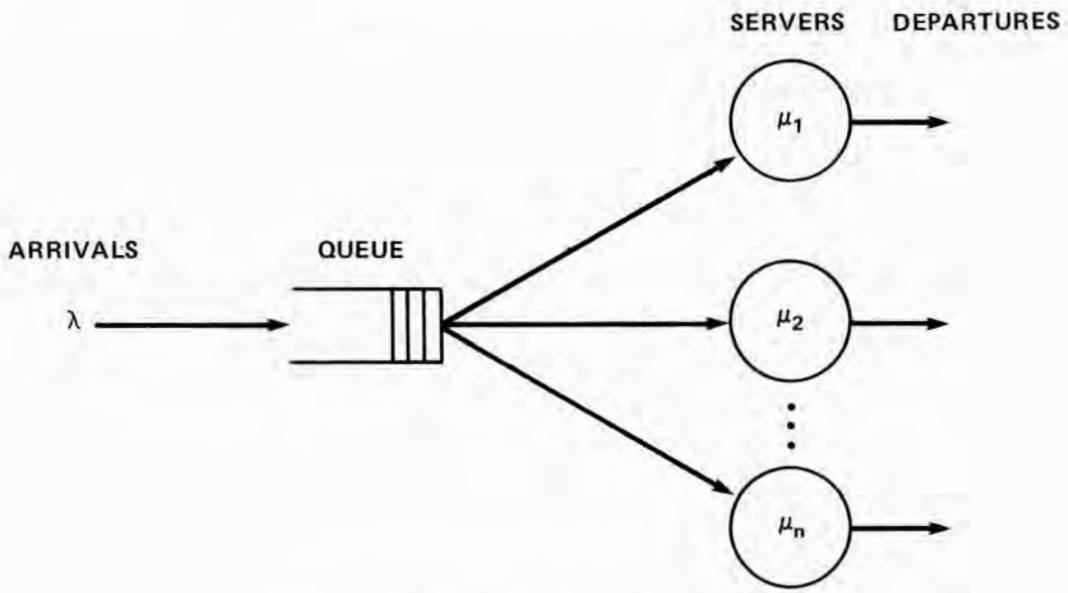



Figure 3-1. Multiple server configuration

The objective of this chapter is to explore several non-preemptive server scheduling disciplines for this system and to develop optimal disciplines to minimize the mean time a customer spends in the system.

## 3.1 General Formulation of Control Model

The system of figure 3-1 can be characterized as a continuous time Markov process and the control of the system can be addressed through the notion of Markovian decision

processes [DERMC70]. Within this context, the state of the system at any instant can be characterized by the number of customers waiting in the queue plus a representation of the busy/idle status of each server. The following state description is introduced to capture this notion. A state $s \epsilon S = (q, n_1 n_2 \ldots n_n)$, where q is the number of customers in the queue, and $n_i$ is zero if server i is idle and one if server i is busy ($1 \leq i \leq n$). This forms a countably infinite state space S. The state of the system at time t is given by $s(t) = (q(t), n(t))$, where $n(t) = n_1(t)\, n_2(t) \ldots n_n(t)$. The total number of customers in the system at time t is then:

$$N(t) = q(t) + \sum_{i=1}^{n} n_i(t) \qquad (1)$$

The instants of transition among states of S are denoted $\tau_i$, $i \geq 1$. The initial state of the system is denoted $s_0$, and subsequent states are denoted $s_i \equiv s(t \mid \tau_i \leq t < \tau_{i+1})$. The sequence $\{s_0, s_1, \ldots\}$ comprises a discrete Markov process. Since the underlying process is a continuous time process, however, admissible state transitions will be limited to singular changes in the state of the system (e.g., arrival of one customer). No bulk arrivals will be considered.

Each point $\tau_i$ is a decision point which influences the behavior of the system during and following the interval $(\tau_i, \tau_{i+1})$. The action taken in this interval is denoted $d_i \epsilon A$, the set of admissible actions. The set A is defined as the set of 2 X n-dimensional matrices [a]. The first row $(\vec{a_1})$ is an n-dimensional vector whose components $a_{1i}$ are all zero with the exception of the $j^{th}$ which is in the interval [0, 1]. This specifies the probability with which the next customer arrival will result in server j being started up. One could conceive of disciplines with more than one component of $\vec{a_1}$ being non-zero to capture notions of possibly starting up more than one server (which violates the singular action assumption) or of probabilistically selecting one among several servers to start up. If the servers being selected have equal service rates, then the two alternatives are equivalent, and if they have dissimilar service rates, then an ordering can be imposed on

server start up and this ordering used to guide the selection of the next server to be initiated. The result is that only one component of $\vec{a_1}$ need be non-zero. The second row of [a], denoted $\vec{a_2}$ is an n-dimensional vector, the $j^{th}$ component of which specifies the probability with which server j will begin service on a new customer (given that one is waiting) following its current service period (given that it is presently busy).

Given the specified state and action spaces, the process evolves according to the transition probability rates $\pi(j|i, [a]), i, j \epsilon S, [a] \epsilon A$, which means that the rate of transition to state j given the system is in state i and action [a] is taken is $\pi(j|i, [a])$. Consideration here is restricted to stationary processes, in which the $\pi(j|i, [a])$ are a function only of the current state of the system, and not of time. We wish to find optimal policies, as reflected in the $\pi(j|i, [a])$ for controlling the system of figure 3-1, where, by policy, we mean an assignment of an action matrix $[a] \epsilon A$ to each state $i \epsilon S$. Optimality is defined with respect to a specific objective function, which, for the greater part of this dissertation will be the mean number of customers in the system.

## 3.2 General Policies for Two Servers

To begin the analysis, we consider the set of stationary, non-preemptive control policies with singular actions. This is illustrated in the state diagram of figure 3.2-1, in which every physically realizable state transition is represented as a directed arc between two states. A specific control policy can be realized by assigning transition rates to each directed arc. Consider state (2, 01), for example, in which two customers are waiting in the queue, server 1 is idle, and server 2 is busy. If a new customer arrives, the system can move to either state (2, 11) or (3, 01), depending on whether or not server 1 is started up. If server 2 completes service on its current customer, the system can move to either state (1, 01) or (2, 00), depending on whether or not server 2 removes the next customer from the queue or becomes idle. Note that a transition from (2, 01) to (1, 10) is not included since it violates the singular action assumption. The effect can be realized, however,

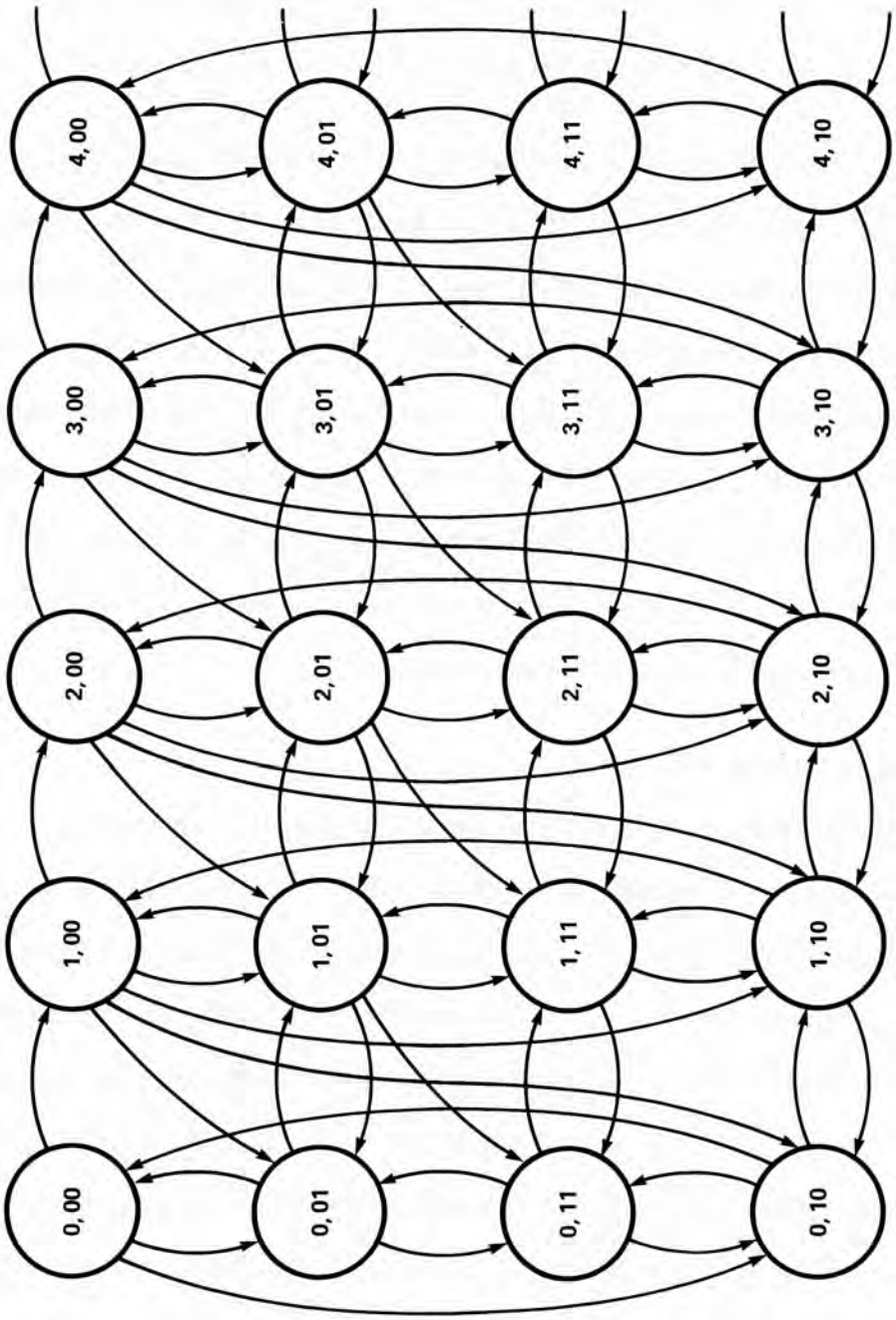



Figure 3.2-1. Admissible State Transitions

through the state sequence (2, 01) → (2, 00) → (1, 10). The final transition out of (2, 01) is to (1, 11), in which server 1 is started up without the occurrence of either an arrival or a departure. No other transitions out of state (2, 01) are physically realizable.

The series of states (i, 00), $i \geq 1$ across the top of the figure correspond to queueing up customers and leaving both servers idle. It can be reasonably argued that if one's objective is to minimize the mean time spent by a customer in the system, then any policy which uses the servers will be an improvement over those policies which use neither server. Similarly, it can be argued that the (i, 01) → (i–1, 11) transitions are inferior to the (i–1, 01) → (i–1, 11) and (i, 11) → (i–1, 11) transitions for $i \geq 1$. The (i, 01) → (i–1, 11) transition occurs between customer arrivals and departures. If starting up the new server improves the overall performance of the system, then that performance improvement can be realized sooner through the (i–1, 01) → (i–1, 11) and (i, 11) → (i–1, 11) transitions rather than using (i, 01) as an intermediate point.

Paring figure 3.2–1 of these sub–optimal states and transitions, one gets figure 3.2–2, the feasible state transitions. Added to each arc is a transition rate $\pi(j|i, [a])$ as described in section 3.1. The action space A is summarized in figure 3.2–3. Noting that $\vec{a_1}$ for state (0, 00) violates the definition given in section 3.1, this case represents a special situation. Given the system is in state (0, 00), when the next customer arrives, a decision must be made as to which server will become busy. As shown, with probability $\eta$, server 1 will be selected, and with probability $(1 - \eta)$, server 2 will be activated. Recalling, again, that the overall objective is to minimize the mean time a customer spends in the system, it is clear that the faster of the two servers should be selected first, thus setting $\eta$ to either 0 or 1. Without loss of generality, we will subsequently consider $\eta$ to be 1, i.e., server 1 will be considered to be the faster of the two servers, and the one to be preferenced given a choice. This implies that if there exists a customer in the queue, then the faster server should never be allowed to stand idle (this is treated rigorously in chapter 4). Hence,

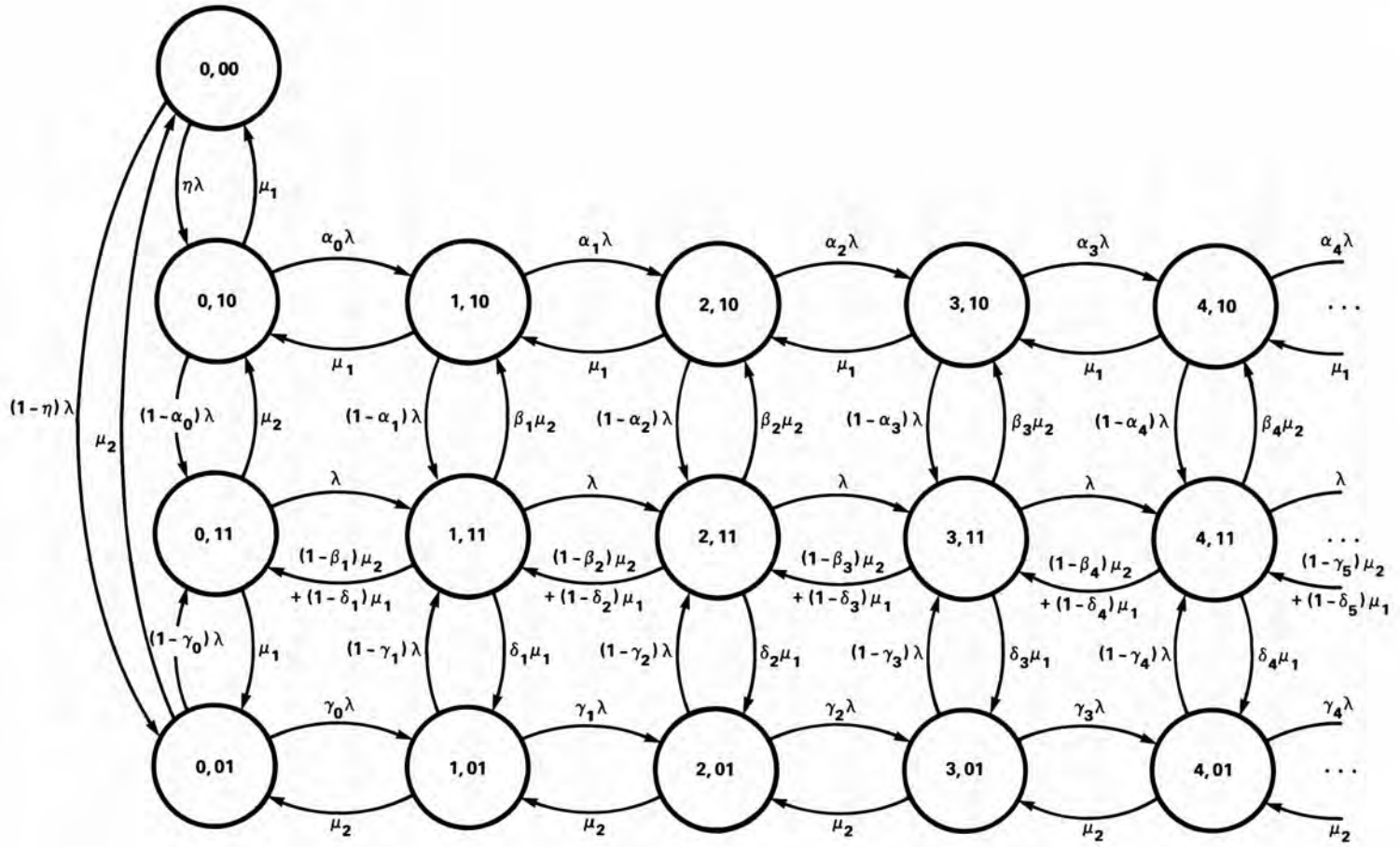



Figure 3.2–2. Feasible State Transitions

| State | [a] |
|---|---|
| (0, 00) | $\begin{bmatrix} \eta, 1-\eta \\ 0, 0 \end{bmatrix}$ |
| (0, 10) | $\begin{bmatrix} 0, 1-\alpha_0 \\ 0, 0 \end{bmatrix}$ |
| (i, 10) $i \geq 1$ | $\begin{bmatrix} 0, 1-\alpha_i \\ 1, 0 \end{bmatrix}$ |
| (0, 11) | $\begin{bmatrix} 0, 0 \\ 0, 0 \end{bmatrix}$ |
| (i, 11) $i \geq 1$ | $\begin{bmatrix} 0, 0 \\ 1-\beta_i, 1-\delta_i \end{bmatrix}$ |
| (0, 01) | $\begin{bmatrix} 1-\gamma_0, 0 \\ 0, 0 \end{bmatrix}$ |
| (i, 01) $i \geq 1$ | $\begin{bmatrix} 1-\gamma_i, 0 \\ 0, 1 \end{bmatrix}$ |

Figure 3.2–3. Action space for figure 3.2–2

figure 3.2–2 can be pared further by setting all $\delta_i$ and $\gamma_i$ to zero, which effectively eliminates all states (i, 01), $i \geq 1$, from the figure, as shown in figure 3.2–4. Figure 3.2–5 shows the revised action space. Now the problem of finding an optimal policy is reduced to one of finding the values of $\alpha_i$, $i \geq 0$ and $\beta_i$, $i \geq 1$ which optimize performance.

The problem to be addressed in this dissertation is a restricted version of the above problem in which:

a)
$$\alpha_i = \begin{cases} 1 \text{ for } 0 \leq i < M-1 \\ \alpha \text{ for } i = M-1 \\ 0 \text{ for } i > M-1 \end{cases} \quad (1)$$

b)
$$\beta_i = \begin{cases} 1 \text{ for } 1 \leq i < M \\ \beta \text{ for } i = M \\ 0 \text{ for } i > M. \end{cases}$$

The state transition diagram for this problem is shown in figure 3.2–6. The discipline so described is called Threshold Queueing, since in operation it begins by using the faster of

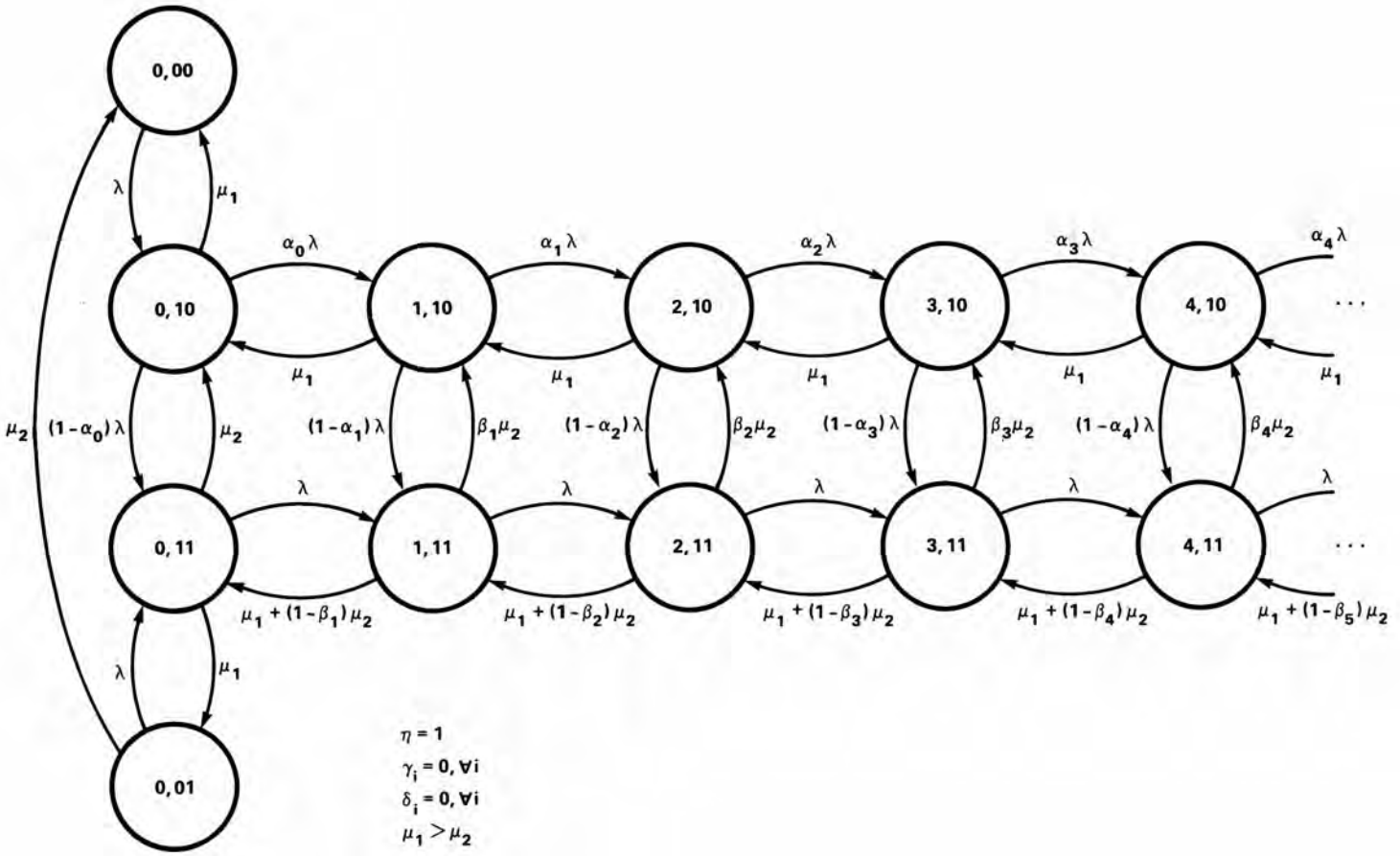



Figure 3.2–4. Feasible State Transitions Preferencing Faster Server

| State | [a] |
|---|---|
| (0, 00) | $\begin{bmatrix} 1, 0 \\ 0, 0 \end{bmatrix}$ |
| (0, 10) | $\begin{bmatrix} 0, 1 - \alpha_0 \\ 0, 0 \end{bmatrix}$ |
| (i, 10) $i \geq 1$ | $\begin{bmatrix} 0, 1 - \alpha_i \\ 1, 0 \end{bmatrix}$ |
| (0, 11) | $\begin{bmatrix} 0, 0 \\ 0, 0 \end{bmatrix}$ |
| (i, 11) $i \geq 1$ | $\begin{bmatrix} 0, 0 \\ 1 - \beta_i, 1 \end{bmatrix}$ |
| (0, 01) | $\begin{bmatrix} 1, 0 \\ 0, 0 \end{bmatrix}$ |

Figure 3.2–5. Action space for figure 3.2–4

the two servers and allows a queue to form while the slower server remains idle. When the queue size builds up to a predetermined threshold value of M–1, the slower server is selected according to a probabilistic rule. When the queue size builds up to M and larger, the slower server is brought into full service until, once again, the queue size falls below M. We proceed in section 3.3 to analyze this discipline.

## 3.3 Threshold Queueing for Two Servers

Markov chain analysis provides a convenient means to analyze the steady state performance of the Threshold Queueing discipline [KLEIL75]. The state transition rate diagram is shown in figure 3.2–6. For a system in steady state, the effective rate of departure from any given state must equal the effective rate of entry. One can utilize figure 3.2–6 to construct the steady state equations by inspection. As an example, the effective rate of departure from state (0, 00) is the rate of departure from that state given the system is in that state multiplied by the probability that the system is in that state. Define $p_{q,n}$ as the probability that the system is in state (q, n). Then the effective rate of departure from state (0, 00) is $\lambda p_{0,00}$. As can be seen from figure 3.2–6,

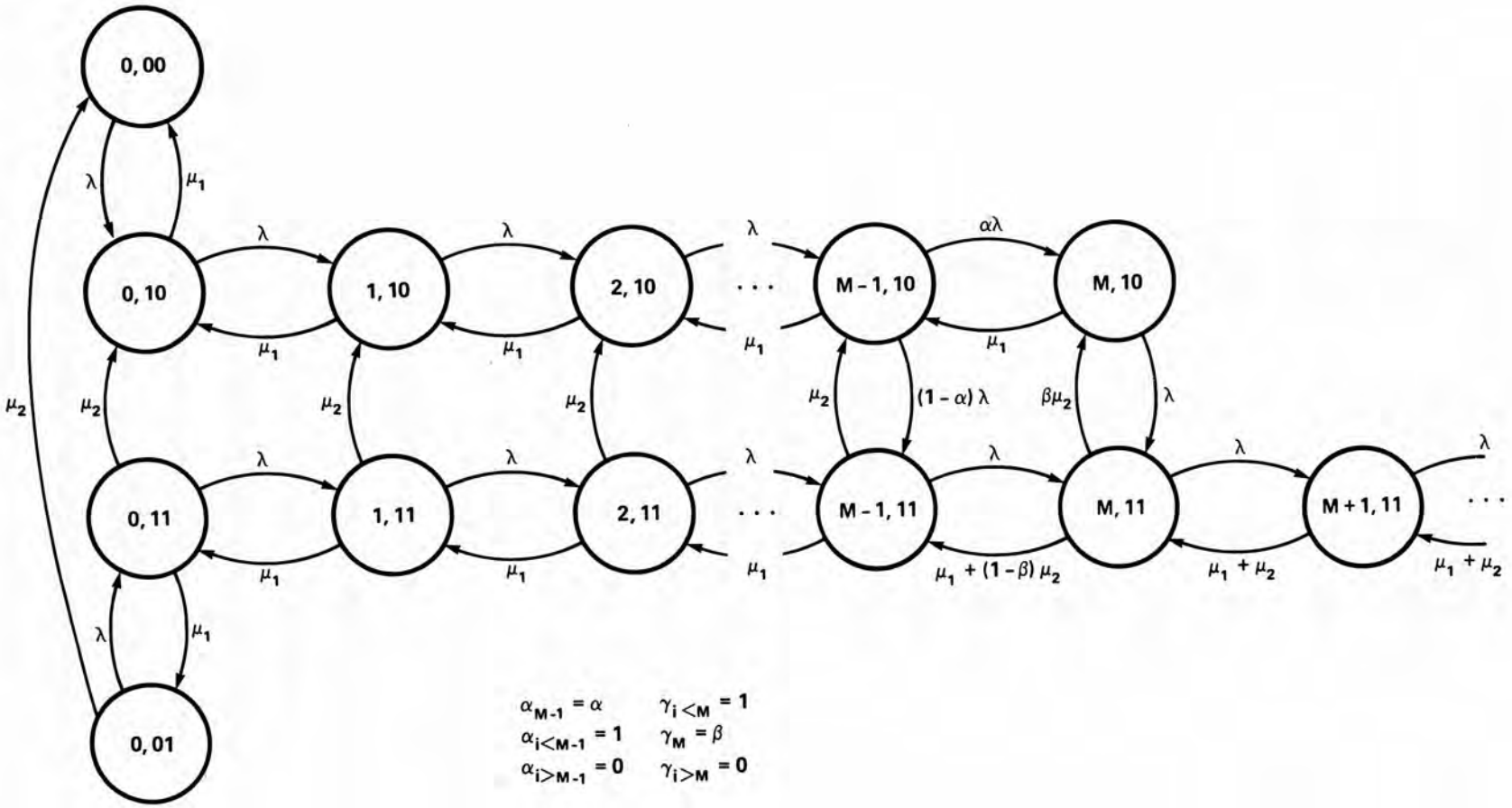



Figure 3.2–6. Threshold Queueing State Transition Diagram

state (0, 00) can be entered from either state (0, 01) at a rate of $\mu_1$, or from state (0, 10) at a rate of $\mu_2$. The effective rate of entry into state (0, 00) is, therefore, $\mu_1 p_{0,01} + \mu_2 p_{0,10}$. This leads directly to the first steady state equation (1a). Proceeding in this manner throughout figure 3.2–6 results in the set of steady state equations (1) which the system must satisfy.

$$\text{a)}\ \lambda p_{0,00} = \mu_1 p_{0,10} + \mu_2 p_{0,01} \tag{1}$$

$$\text{b)}\ (\lambda + \mu_1)\, p_{0,10} = \lambda p_{0,00} + \mu_1 p_{1,10} + \mu_2 p_{0,11}$$

$$\text{c)}\ (\lambda + \mu_1)\, p_{i,10} = \lambda p_{i-1,10} + \mu_1 p_{i+1,10} + \mu_2 p_{i,11} \quad 1 \le i \le M - 1$$

$$\text{d)}\ (\lambda + \mu_2)\, p_{0,01} = \mu_1 p_{0,11}$$

$$\text{e)}\ (\lambda + \mu_1 + \mu_2)\, p_{0,11} = \lambda p_{0,01} + \mu_1 p_{1,11}$$

$$\text{f)}\ (\lambda + \mu_1 + \mu_2)\, p_{i,11} = \lambda p_{i-1,11} + \mu_1 p_{i+1,11} \quad 1 \le i \le M - 2$$

$$\text{g)}\ (\lambda + \mu_1 + \mu_2)\, p_{M-1,11} = \lambda p_{M-2,11} + (1 - \alpha)\lambda p_{M-1,10} + [\mu_1 + (1 - \beta)\mu_2]\, p_{M,11}$$

$$\text{h)}\ (\lambda + \mu_1)\, p_{M,10} = \alpha\lambda p_{M-1,10} + \beta\mu_2 p_{M,11}$$

$$\text{i)}\ (\lambda + \mu_1 + \mu_2)\, p_{M,11} = \lambda p_{M-1,11} + \lambda p_{M,10} + (\mu_1 + \mu_2)\, p_{M+1,11}$$

$$\text{j)}\ (\lambda + \mu_1 + \mu_2)\, p_{M+i,11} = \lambda p_{M+i-1,11} + (\mu_1 + \mu_2)\, p_{M+i+1,11} \quad i \ge 1$$

Eqn (1) can be expressed in terms of two normalized variables $v_1$ and $v_2$ defined as:

$$\text{a)}\ v_1 = \mu_1/\lambda \tag{2}$$

$$\text{b)}\ v_2 = \mu_2/\lambda$$

The revised equations then become:

$$\text{a)}\ p_{0,00} = v_1 p_{0,10} + v_2 p_{0,01} \tag{3}$$

$$\text{b)}\ (1 + v_1)\, p_{0,10} = p_{0,00} + v_1 p_{1,10} + v_2 p_{0,11}$$

$$\text{c)}\ (1 + v_1)\, p_{i,10} = p_{i-1,10} + v_1 p_{i+1,10} + v_2 p_{i,11} \quad 1 \le i \le M - 1$$

$$\text{d)}\ (1 + v_2)\, p_{0,01} = v_1 p_{0,11}$$

$$\text{e)}\ (1 + v_1 + v_2)\, p_{0,11} = p_{0,01} + v_1 p_{1,11}$$

$$\text{f)}\ (1 + v_1 + v_2)\, p_{i,11} = p_{i-1,11} + v_1 p_{i+1,11} \quad 1 \le i \le M - 2$$

$$\text{g)}\ (1 + v_1 + v_2)\, p_{M-1,11} = p_{M-2,11} + (1 - \alpha)\, p_{M-1,10} + [v_1 + (1 - \beta)\, v_2]\, p_{M,11}$$

$$\text{h)}\ (1 + v_1)\, p_{M,10} = \alpha p_{M-1,10} + \beta v_2 p_{M,11}$$

i) $(1 + v_1 + v_2)\, p_{M,11} = p_{M-1,11} + p_{M,10} + (v_1 + v_2)\, p_{M+1,11}$

j) $(1 + v_1 + v_2)\, p_{M+i,11} = p_{M+i-1,11} + (v_1 + v_2)\, p_{M+i+1,11} \qquad i \geq 1$

Equations (3a–j) form an infinite set of linear equations for the p's. This set can be reduced to a finite set by summing the state probabilities for all states (i, 11) for $i > M$ into an equivalent single state.* Denoting this state ($> M$, 11), its steady state probability is given by

$$p_{>M,11} = \sum_{i=1}^{\infty} p_{M+i,11} = \left(\frac{1}{v_2 + v_1 - 1}\right) p_{M,11}. \tag{4}$$

Equation (4) arises from noting that beyond state (M, 11) the system performs as an M/M/1 system with service rate $\mu_1 + \mu_2$. Equations (3a–i) plus eqn (3) can now be solved directly to yield the following algorithmic solution for the p's:

a) $c_{0,01} = 1$ b) $c_{0,11} = \dfrac{v_2 + 1}{v_1}$ (5)

c) $c_{1,11} = \left(\dfrac{v_2 + 1}{v_1}\right)^2 + \dfrac{v_2}{v_1}$

d) $c_{i,11} = \left(\dfrac{v_2 + v_1 + 1}{v_1}\right) c_{i-1,11} - \dfrac{1}{v_1}\, c_{i-2,11} \qquad 2 \leq i \leq M - 1$

e) $c_{0,00} = v_1^M \left[1 + \dfrac{v_2}{v_1}(1 - \beta) + (1 - \alpha)(v_2 + v_1)\right]^{-1} \left\{\beta v_2 \left[(v_2 + v_1 + 1)\, c_{M-1,11} - c_{M-2,11}\right]\right.$

$$+ (1 + v_1)\left[v_1 + (1 - \beta)\, v_2\right] \frac{v_2}{v_1}\left[\sum_{j=0}^{M}\left(\frac{1}{v_1}\right)^j + \sum_{k=0}^{M-1}\sum_{j=0}^{k}\left(\frac{1}{v_1}\right)^j c_{M-k-1,11}\right]$$

$$\left. - \left[\alpha v_1 + (\alpha - \beta)\, v_2\right] \frac{v_2}{v_1}\left[\sum_{j=0}^{M-1}\left(\frac{1}{v_1}\right)^j + \sum_{k=0}^{M-2}\sum_{j=0}^{k}\left(\frac{1}{v_1}\right)^j c_{M-k-2,11}\right]\right\}$$

f) $c_{0,10} = \dfrac{1}{v_1}(c_{0,00} - v_2)$ g) $c_{i,10} = \dfrac{1}{v_1}\left(c_{i-1,10} - v_2 - v_2 \sum_{j=0}^{i-1} c_{j,11}\right) \qquad 1 \leq i \leq M$

h) $c_{M,11} = \dfrac{1}{\beta v_2}\left[(v_1 + 1)\, c_{M,10} - \alpha c_{M-1,10}\right]$ i) $c_{>M,11} = \left(\dfrac{1}{v_2 + v_1 - 1}\right) c_{M,11}$

*Replacement of states (i, 11) for $i > M$ with an equivalent state such that the steady state probabilities remain unchanged is done to reduce the mathematical complexity of the problem without changing its probabilistic structure. In computing performance parameters such as the mean number of customers in the system, the complete structure of the state transition diagram must be considered. The presence of simple relationships among the state probabilities for the infinite Markov chain makes this computation straightforward. The details are contained in Appendix C.

j) $$p_{0,01} = \left( c_{0,00} + c_{0,01} + \sum_{i=0}^{M} c_{i,10} + \sum_{i=0}^{M-1} c_{i,11} + c_{M,11} + c_{>M,11} \right)^{-1}$$

k) $p_{q,n} = c_{q,n}\, p_{0,01}$ for all states (q, n)

Appendix A contains the complete derivation of equations (5a–k). These equations can be solved further to get a non-algorithmic closed form expression for each $p_{q,n}$. This form of the solution is derived in Appendix B and summarized here. We first introduce two sets of functions f and g:

$$f_i(x, y) = \sum_{j=0}^{\left\lfloor \frac{i}{2} \right\rfloor} (-1)^j \binom{i-j}{j} x^{i-2j}\, y^j \tag{6}$$

$$g_k(x, y) = \sum_{i=0}^{k} y^{k-i}\, f_i(x, y) \tag{7}$$

For notational convenience, $f_i$ and $g_k$ will be used subsequently to refer to $f_i\left(\frac{v_2 + v_1 + 1}{v_1}, \frac{1}{v_1}\right)$ and $g_k\left(\frac{v_2 + v_1 + 1}{v_1}, \frac{1}{v_1}\right)$, respectively. The solution then becomes:

a) $$c_{i,11} = f_{i+1} - f_i \qquad 0 \le i \le M - 1 \tag{8}$$

b) $$c_{0,00} = \frac{v_1^M v_2 \left\{ -(v_2 + v_1)\, g_{M-1}\, \alpha + [v_1 (f_M - f_{M-1}) - v_2 g_M]\, \beta + (v_1 + 1)(v_2 + v_1)\, g_M \right\}}{-v_1 (v_2 + v_1)\, \alpha - v_2 \beta + (v_1 + 1)(v_2 + v_1)}$$

c) $$c_{i,10} = \frac{v_2 (v_2 + v_1)\left(g_i - v_1^{M-1-i}\, g_{M-1}\right) \alpha + \left[ v_1^{M-1-i}\, v_2 (v_1 f_M - v_1 f_{M-1} - v_2 g_M) + \frac{v_2^2}{v_1} g_i \right] \beta + v_2 (v_1 + 1)(v_2 + v_1) \left( v_1^{M-1-i}\, g_M - \frac{1}{v_1} g_i \right)}{-v_1 (v_2 + v_1)\, \alpha - v_2 \beta + (v_1 + 1)(v_2 + v_1)} \qquad 0 \le i \le M$$

d) $$c_{M,11} = \frac{-[(v_1 - v_2) f_M - v_1 f_{M-1}]\, \alpha + (v_1 + 1)(f_M - f_{M-1})}{-v_1 (v_2 + v_1)\, \alpha - v_2 \beta + (v_1 + 1)(v_2 + v_1)}$$

e) $$c_{M+i,11} = \left( \frac{1}{v_2 + v_1} \right)^i c_{M,11}$$

$$\text{f)}\ \sum_{i=0}^{\infty} c_{M+i,11} = \left(\frac{v_2 + v_1}{v_2 + v_1 - 1}\right) c_{M,11}$$

$$\left\{-(v_2 + v_1)\left[v_2(v_2 + v_1 - 1)\sum_{i=0}^{M-1}\sum_{j=0}^{i} v_1^{j+1} f_i + (v_1^2 - v_2^2) f_M - v_1 f_{M-1}\right]\alpha\right.$$

$$-v_2(v_2 + v_1 - 1)\left[v_2\sum_{i=0}^{M}\sum_{j=0}^{i} v_i^j f_i + f_{M-1}\sum_{i=0}^{M+1} v_1^i - f_M\sum_{i=1}^{M+1} v_1^i\right]\beta$$

$$\left.+(v_1 + 1)(v_2 + v_1)\left[(v_2 + v_1 - 1)\left(f_M + v_2\sum_{i=0}^{M}\sum_{j=0}^{i} v_1^j f_i\right) + f_M - f_{M-1}\right]\right\}$$

$$\text{g)}\ p_{0,01}^{-1} = \frac{}{(v_2 + v_1 - 1)\left[-v_1(v_2 + v_1)\alpha - v_2\beta + (v_1 + 1)(v_2 + v_1)\right]}$$

Given the steady state probabilities of state occupancy, one can derive expressions for the expected utilization of each server ($\rho_i$):

$$\text{a)}\ \rho_1 = \sum_{i=0}^{M} p_{i,10} + \sum_{i=0}^{\infty} p_{i,11} = 1 - p_{0,00} - p_{0,01} \tag{9}$$

$$\text{b)}\ \rho_2 = p_{0,01} + \sum_{i=0}^{\infty} p_{i,11} = p_{0,01} + \sum_{i=0}^{M-1} p_{i,11} + \left(\frac{v_2 + v_1}{v_2 + v_1 - 1}\right) p_{M,11}$$

Substituting for the $p_{q,n}$ and simplifying yields the following:

$$\left\{-(v_2 + v_1)\left[v_2(v_2 + v_1 - 1)\sum_{i=1}^{M-1}\sum_{j=1}^{i} v_1^j f_i + (v_1^2 - v_2^2) f_M - v_1 f_{M-1} - (v_2 + v_1 - 1) v_1\right]\alpha\right.$$

$$-v_2(v_2 + v_1 - 1)\left[v_2\sum_{i=1}^{M}\sum_{j=1}^{i} v_1^{j-1} f_i + f_{M-1}\sum_{i=0}^{M} v_1^i - f_M\sum_{i=1}^{M} v_1^i - 1\right]\beta$$

$$\left.+(v_1 + 1)(v_2 + v_1)\left[(v_2 + v_1 - 1)\left(f_M - 1 + v_2\sum_{i=1}^{M}\sum_{j=1}^{i} v_1^{j-1} f_i\right) + f_M - f_{M-1}\right]\right\}$$

$$\text{a)}\ \rho_1 = \frac{}{\left\{-(v_2 + v_1)\left[v_2(v_2 + v_1 - 1)\sum_{i=0}^{M-1}\sum_{j=0}^{i} v_1^{j+1} f_i + (v_1^2 - v_2^2) f_M - v_1 f_{M-1}\right]\alpha\right.} \tag{10}$$

$$-v_2(v_2 + v_1 - 1)\left[v_2\sum_{i=0}^{M}\sum_{j=0}^{i} v_1^j f_i + f_{M-1}\sum_{i=0}^{M+1} v_1^i - f_M\sum_{i=1}^{M+1} v_1^i\right]\beta$$

$$\left.+(v_1 + 1)(v_2 + v_1)\left[(v_2 + v_1 - 1)\left(f_M + v_2\sum_{i=0}^{M}\sum_{j=0}^{i} v_1^j f_i\right) + f_M - f_{M-1}\right]\right\}$$

$$\{(v_2 + v_1)\,[(v_2 - v_1 v_2 - v_1^2)\,f_M + v_1 f_{M-1}]\,\alpha - v_2(v_2 + v_1 - 1)\,f_M \beta$$

$$+(v_1 + 1)\,(v_2 + v_1)\,[(v_2 + v_1)\,f_M - f_{M-1}]\}$$

b) $$\rho_2 = \frac{}{\left\{-(v_2 + v_1)\left[v_2(v_2 + v_1 - 1)\sum_{i=0}^{M-1}\sum_{j=0}^{i} v_1^{j+1} f_i + (v_1^2 - v_2^2)\,f_M - v_1 f_{M-1}\right]\alpha\right.}$$

$$-v_2(v_2 + v_1 - 1)\left[v_2 \sum_{i=0}^{M}\sum_{j=0}^{i} v_1^{j} f_i + f_{M-1}\sum_{i=0}^{M+1} v_1^{i} - f_M \sum_{i=1}^{M+1} v_1^{i}\right]\beta$$

$$\left.+(v_1 + 1)\,(v_2 + v_1)\left[(v_2 + v_1 - 1)\left(f_M + v_2 \sum_{i=0}^{M}\sum_{j=0}^{i} v_1^{j} f_i\right) + f_M - f_{M-1}\right]\right\}$$

These equations comprise the solution to the queueing system of figure 3.2–6. The fundamental solution expresses the steady state probabilities associated with each system state. Other performance measures such as server utilization are derived from these probabilities. Note that the action space of figure 3.2–5 with $\alpha_i = 1$ for $0 \leq i < M - 1$, $\alpha_i = 0$ for $i \geq M$, $\beta_i = 1$ for $1 \leq i < M$, and $\beta_i = 0$ for $i \geq M + 1$ is assumed given. The problem of optimization will be considered next.

# CHAPTER 4

## OPTIMIZATION OF THRESHOLD QUEUEING FOR TWO SERVERS

The primary performance criterion which will be used to optimize the performance of threshold queueing will be the mean number of customers in the system ($\bar{N}$). Little's result [LITTJ61] provides the very straightforward relationship between this and the mean time a customer spends in the system ($\bar{T}$):

$$\bar{T} = \bar{N}/\lambda \tag{1}$$

The overall objective will be to minimize $\bar{N}$, and consequently $\bar{T}$. The mean queue length $\bar{Q}$ will be of interest also. Given the steady state probabilities (eqn (3.3-8)) associated with each state, we have the following relationships (recalling eqn (3.3-9)):

$$\text{a) } \bar{Q} = \sum_{i=1}^{M} i p_{i,10} + \sum_{i=1}^{\infty} i p_{i,11} \tag{2}$$

$$\text{b) } \bar{N} = \bar{Q} + \rho_1 + \rho_2$$

$$= p_{0,01} + \sum_{i=0}^{M} (i+1)\, p_{i,10} + \sum_{i=0}^{\infty} (i+2)\, p_{i,11}$$

$$= \left[ 1 + \sum_{i=0}^{M} (i+1)\, c_{i,10} + \sum_{i=0}^{M-1} (i+2)\, c_{i,11} + \sum_{i=0}^{\infty} (M+2+i)\, c_{M+i,11} \right] p_{0,01}$$

Appendix C contains the derivation of $\bar{N}$ which culminates in the following:

$$\bar{N}_M(v_1, v_2) = \frac{H_1^M(v_1, v_2)\,\alpha + H_2^M(v_1, v_2)\,\beta + H_3^M(v_1, v_2)}{(v_2 + v_1 - 1)\,[H_4^M(v_1, v_2)\,\alpha + H_5^M(v_1, v_2)\,\beta + H_6^M(v_1, v_2)]} \tag{3}$$

where

$$\text{a) } H_1^M(v_1, v_2) = -(v_2 + v_1) \left\{ (v_2 + v_1 - 1)^2 \left[ v_2 \sum_{i=1}^{M-1} \sum_{j=i}^{M-1} i v_1^{j-i+1} f_j - v_1 \sum_{i=0}^{M-1} f_i \right] \right. \tag{4}$$

$$\left. + (v_1^2 - v_2^2)\,[(M+1)(v_2 + v_1 - 1) + 1]\, f_M - v_1 [(M+2)(v_2 + v_1 - 1) + 1]\, f_{M-1} \right\}$$

$$\text{b) } H_2^M(v_1, v_2) = -v_2 (v_2 + v_1 - 1)^2 \left\{ v_2 \sum_{i=1}^{M} \sum_{j=i}^{M} i v_1^{j-i} f_j + (f_{M-1} - f_M) \sum_{i=1}^{M} i v_1^{M-i+1} - \sum_{i=0}^{M-2} f_i + M f_{M-1} \right\}$$

c) $H_3^M(v_1, v_2) = (v_1 + 1)(v_2 + v_1)\left\{\left[v_2 \sum_{i=1}^{M}\sum_{j=i}^{M} i v_1^{j-i} f_j - \sum_{i=0}^{M-1} f_i + (M+1) f_M\right](v_2 + v_1 - 1)^2 + [(M+2)(v_2 + v_1 - 1) + 1](f_M - f_{M-1})\right\}$

d) $H_4^M(v_1, v_2) = -(v_2 + v_1)\left[(v_1^2 - v_2^2) f_M - v_1 f_{M-1} + v_2(v_2 + v_1 - 1) \sum_{i=0}^{M-1}\sum_{j=0}^{i} v_1^{j+1} f_i\right]$

e) $H_5^M(v_1, v_2) = -v_2(v_2 + v_1 - 1)\left[v_2 \sum_{i=0}^{M}\sum_{j=0}^{i} v_1^j f_i - f_M \sum_{i=1}^{M+1} v_1^i + f_{M-1} \sum_{i=0}^{M+1} v_1^i\right]$

f) $H_6^M(v_1, v_2) = (v_1 + 1)(v_2 + v_1)\left[\left(v_2 \sum_{i=0}^{M}\sum_{j=0}^{i} v_1^j f_i + f_M\right)(v_2 + v_1 - 1) + f_M - f_{M-1}\right]$

In order to continue further analysis of $\bar{N}$ and the $H_i^M$'s, some properties of $f_i$ and its domain will first be established. The domain of the function $f_i(x, y)$ where $x = \frac{v_1 + v_2 + 1}{v_1}$ and $y = \frac{1}{v_1}$ is determined by the ergodic constraint $v_1 + v_2 > 1$ and the ordering convention $v_1 \geq v_2 > 0$. This domain is illustrated in figure 4–1, and includes the bounds $x > y + 1$ and $0 < y < 2$.

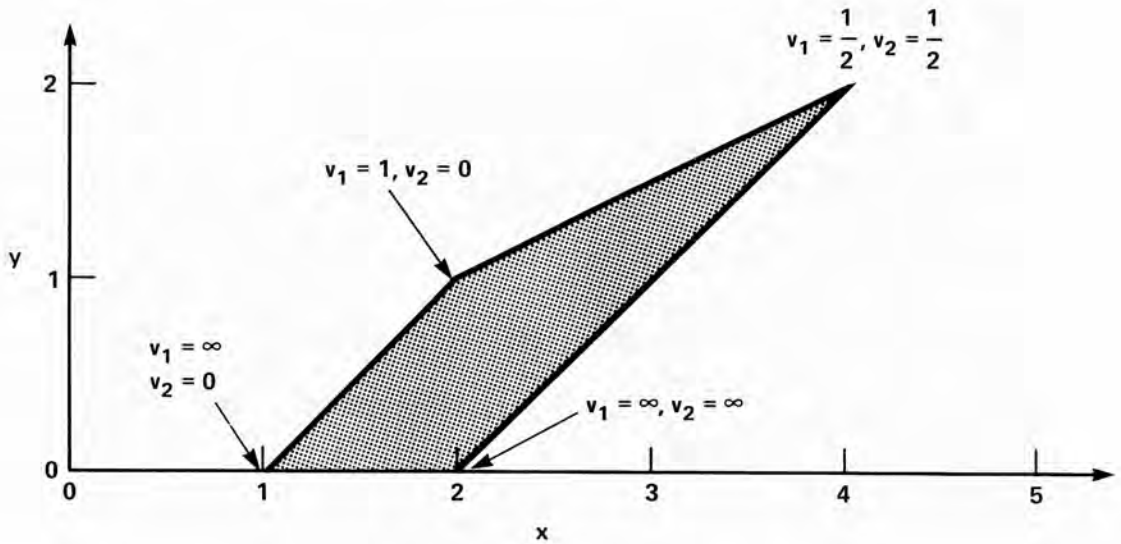



Figure 4–1. Domain of $f_i(x, y)$

The following result is fundamental to the analysis of $\bar{N}$.

Lemma 4–1 The function $f_i(x, y) = \sum_{j=0}^{\lfloor \frac{i}{2} \rfloor} (-1)^j \binom{i-j}{j} x^{i-2j} y^j$ over the domain shown in figure 4–1 is strictly positive and monotonically increasing with respect to i.

Proof The proof is by induction on i. For the initial case, we have:

a) $f_0(x, y) = 1$ (5)

b) $f_1(x, y) = x > f_0(x, y)$

Now, for the inductive step, assume that $f_i(x, y) > f_{i-1}(x, y)$. Utilizing the following identity derived as eqn (6) of Appendix B:

$$f_i(x, y) = xf_{i-1}(x, y) - yf_{i-2}(x, y) \tag{6}$$

for $f_{i+1}(x, y)$ we have:

$$f_{i+1}(x, y) = xf_i(x, y) - yf_{i-1}(x, y) > (y + 1) f_i(x, y) - yf_{i-1}(x, y) \tag{7}$$

from which the result follows:

$$f_{i+1}(x, y) - f_i(x, y) > y[f_i(x, y) - f_{i-1}(x, y)] > 0 \tag{8}$$

which yields the desired result:

$$f_{i+1}(x, y) > f_i(x, y) \tag{9}$$

We continue the analysis by establishing several relationships among the $H_i$'s which will be needed subsequently.

Lemma 4–2 The functions $H_6^M(v_1, v_2)$, $H_4^M(v_1, v_2) + H_6^M(v_1, v_2)$, $H_5^M(v_1, v_2) + H_6^M(v_1, v_2)$ and $H_4^M(v_1, v_2) + H_5^M(v_1, v_2) + H_6^M(v_1, v_2)$ are all strictly positive for an ergodic system $(v_2 + v_1 > 1)$.

Proof Each function will be treated separately.

Case 1: $H_6^M(v_1, v_2) > 0$ by inspection, recalling Lemma 4–1.

Case 2: $H_4^M(v_1, v_2) + H_6^M(v_1, v_2)$ (10)

$$= (v_2 + v_1) \left\{ v_2(v_2 + v_1 - 1) \left[ \sum_{i=0}^{M} \sum_{j=0}^{i} v_1^j f_i + v_1 \sum_{j=0}^{M} v_1^j f_i \right] + v_2(v_2 + v_1) f_M \right.$$
$$\left. + (v_1 + v_2) f_M - f_{M-1} \right\} > 0$$

Case 3: $H_5^M(v_1, v_2) + H_6^M(v_1, v_2)$ (11)

$$= (v_2 + v_1 - 1) \left\{ v_1(v_2 + v_1 + 1) \left[ v_2 \sum_{i=0}^{M} \sum_{j=0}^{i} v_1^j f_i + f_M \right] + v_2(f_M - f_{M-1}) \sum_{i=0}^{M+1} v_1^i \right\}$$
$$+ (v_1 + 1)(v_2 + v_1)(f_M - f_{M-1}) > 0$$

Case 4: $H_4^M(v_1, v_2) + H_5^M(v_1, v_2) + H_6^M(v_1, v_2)$ (12)

$$= v_2(v_2 + v_1 - 1) \left[ v_1(v_2 + v_1) f_M \sum_{i=0}^{M} v_1^i + (f_M - f_{M-1}) \sum_{i=0}^{M+1} v_1^i + v_1 \sum_{i=0}^{M} \sum_{j=0}^{i} v_1^j f_i \right]$$
$$+ (v_2 + v_1)(f_M - f_{M-1}) + [v_2(v_2 + v_1)^2 + v_1(v_2 + v_1 - 1)] f_M > 0$$

4.1 Probabilistic Thresholds are Sub-optimal

Given eqn (4-3) for $\bar{N}$, the major results of this section can now be presented. These are presented in the form of two theorems.

Theorem 4.1-1 A probabilistic decision rule is suboptimal for threshold queueing in a two server system when the objective is to minimize the mean number of customers in the system.

Proof To minimize $\bar{N}$ with respect to $\alpha$ and $\beta$, the following two equations must be solved simultaneously, where $H_i$ denotes $H_i^M(v_1, v_2)$:

$$\text{a)} \quad \frac{\partial \bar{N}}{\partial \alpha} = \frac{(H_1 H_5 - H_2 H_4)\beta + H_1 H_6 - H_3 H_4}{(v_2 + v_1 - 1)(H_4\alpha + H_5\beta + H_6)^2} = 0 \qquad (1)$$

$$\text{b)} \quad \frac{\partial \bar{N}}{\partial \beta} = \frac{(H_2 H_4 - H_1 H_5)\alpha + H_2 H_6 - H_3 H_5}{(v_2 + v_1 - 1)(H_4\alpha + H_5\beta + H_6)^2} = 0$$

While the system remains ergodic ($v_2 + v_1 - 1 > 0$), $\bar{N}$ remains well-behaved in the physically realizable interval [0, 1] for $\alpha$ and $\beta$, and, hence, its partial derivatives are well-behaved. From eqn (1), we have:

$$\text{a)}\ |\alpha| = \frac{|H_2H_6 - H_3H_5|}{|H_2H_4 - H_1H_5|} \qquad \text{b)}\ |\beta| = \frac{|H_1H_6 - H_3H_4|}{|H_2H_4 - H_1H_5|} \tag{2}$$

In Appendix D it is shown that

$$\text{a)}\ H_1H_6 - H_3H_4 = (v_1 + 1)(v_2 + v_1 - 1)(v_2 + v_1)^2 f_M F_M(v_1, v_2) \tag{3}$$

$$\text{b)}\ H_2H_6 - H_3H_5 = (v_1 + 1)(v_2 + v_1 - 1)(v_2 + v_1)(f_M - f_{M-1}) F_M(v_1, v_2)$$

$$\text{c)}\ H_1H_5 - H_2H_4 = (v_2 + v_1 - 1)(v_2 + v_1)[(v_1 - v_2) f_M - v_1 f_{M-1}] F_M(v_1, v_2)$$

where

$$\begin{aligned} F_M(v_1, v_2) = v_2 &\left[ (M+2)(v_2 + v_1 - 1)(v_2 + v_1) + 1 + (v_2 + v_1 - 1)^2 \left( v_2 \sum_{i=1}^{M} j v_1^{M-j} - 1 \right) \right] \sum_{i=0}^{M-1} \sum_{j=0}^{i} v_1^{j+1} f_i \\ &- v_1(v_2 + v_1 - 1) \left[ 1 + (v_2 + v_1 - 1) \left( v_2 \sum_{j=0}^{M} v_1^j + 1 \right) \right] \left( v_2 \sum_{i=1}^{M-1} \sum_{j=i}^{M-1} i v_1^{j-i} f_j - \sum_{i=0}^{M-2} f_i \right) \\ &- v_2(v_1^2 - v_2^2)[(M+1)(v_2 + v_1 - 1) + 1] \sum_{i=0}^{M} \sum_{j=0}^{i} v_1^j f_i \\ &+ (v_1^2 - v_2^2)(v_2 + v_1 - 1) \left( v_2 \sum_{i=1}^{M} \sum_{j=i}^{M} i v_1^{j-i} f_j - \sum_{i=0}^{M-1} f_i \right) \\ &+ \left\{ v_1 v_2 [(v_2 + v_1)^2 + M(v_2 + v_1 - 1)] \sum_{j=0}^{M} v_1^j - v_1 v_2 (v_2 + v_1 - 1) \sum_{j=1}^{M} j v_1^{M-j} \right. \\ &\left. + v_1(v_2 + v_1)^2 - v_1^2 + v_2^2 \right\} f_{M-1} \end{aligned} \tag{4}$$

so that eqn (2) can be written:

$$\text{a)}\ |\alpha| = \frac{(v_1 + 1)(f_M - f_{M-1})}{|(v_1 - v_2) f_M - v_1 f_{M-1}|} \qquad \text{b)}\ |\beta| = \frac{(v_1 + 1)(v_2 + v_1) f_M}{|(v_1 - v_2) f_M - v_1 f_{M-1}|} \tag{5}$$

It will now be shown that eqn (5) yields $|\alpha| > 1$ and $|\beta| > 1$, i.e.,

$$\text{a)}\ (v_1 + 1)(f_M - f_{M-1}) > |(v_1 - v_2) f_M - v_1 f_{M-1}| \tag{6}$$

$$\text{b)}\ (v_1 + 1)(v_2 + v_1) f_M > |(v_1 - v_2) f_M - v_1 f_{M-1}|$$

Examining (6a) first, it implies two inequalities:

$$\text{a) } (v_2 + 1)f_M - f_{M-1} > 0 \qquad \text{b) } (2v_1 + 1)(f_M - f_{M-1}) - v_2 f_M > 0 \tag{7}$$

While (7a) is true by inspection, (7b) requires one more step. Recalling that $v_1 \geq v_2$, we can write:

$$(2v_1 + 1)(f_M - f_{M-1}) - v_2 f_M \geq (v_2 + v_1 + 1)(f_M - f_{M-1}) - v_2 f_M = f_M - f_{M-2} > 0 \tag{8}$$

where the identity

$$v_1 f_i = (v_2 + v_1 + 1) f_{i-1} - f_{i-2} \tag{9}$$

was employed. Turning to (6b), the following two inequalities result:

$$\text{a) } v_1(v_2 + v_1) f_M + 2v_2 f_M + v_1 f_{M-1} > 0 \tag{10}$$

$$\text{b) } v_1(v_2 + v_1) f_M + v_1(2f_M - f_{M-1}) > 0$$

Both of these are true by inspection, confirming that the solution to eqn (1) yields $|\alpha| > 1$ and $|\beta| > 1$. Since $\alpha$ and $\beta$ are constrained to the interval [0, 1], and $\bar{N}$ is continuous with respect to $\alpha$ and $\beta$, the minimum $\bar{N}$ occurs at one of the points $(\alpha, \beta) = (0, 0), (0, 1), (1, 0)$, or (1, 1), as shown in figure 4.1-1. This completes the proof of theorem 4.1-1.

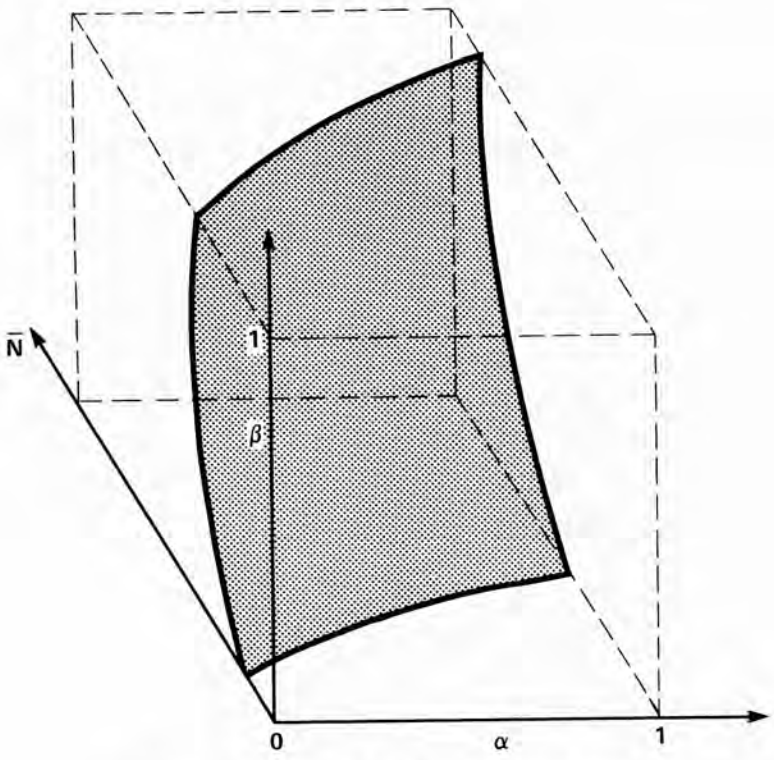



Figure 4.1-1. Probabilistic control surface

Depicted in the figure is the case where $\bar{N}(\alpha, \beta) = \bar{N}(0, 0) \leq \bar{N}(1, 0) \leq \bar{N}(1, 1)$ and $\bar{N}(0, 0) \leq \bar{N}(0, 1) \leq \bar{N}(1, 1)$. In this case it is clear that $(\alpha, \beta) = (0, 0)$ minimizes $\bar{N}$. This leads to the second major result.

Theorem 4.1-2 The minimum of $\bar{N}$ with respect to $\alpha$ and $\beta$ occurs at either the point $(\alpha, \beta) = (0, 0)$ or $(\alpha, \beta) = (1, 1)$.

Proof To prove this theorem, it suffices to prove the following two cases:

Case 1. $(\alpha, \beta) = (0, 0)$: If $\bar{N}(1, 0) > \bar{N}(0, 0)$ then $\bar{N}(0, 1) > \bar{N}(0, 0)$, $\bar{N}(1, 1) > \bar{N}(0, 1)$, and $\bar{N}(1, 1) > \bar{N}(1, 0)$.

Case 2. $(\alpha, \beta) = (1, 1)$: If $\bar{N}(1, 0) < \bar{N}(0, 0)$ then $\bar{N}(0, 1) < \bar{N}(0, 0)$, $\bar{N}(1, 1) < \bar{N}(0, 1)$, and $\bar{N}(1, 1) < \bar{N}(1, 0)$.

Consider case 1.

$$\bar{N}(1, 0) = \frac{H_1 + H_3}{H_4 + H_6} > \bar{N}(0, 0) = \frac{H_3}{H_6} \tag{11}$$

Since $H_6$ and $H_4 + H_6$ are strictly positive for $v_2 + v_1 > 1$ (Lemma 4-2), then inequality (11) is equivalent to:

$$H_1 H_6 > H_3 H_4. \tag{12}$$

To prove the theorem for case 1, it must be shown that if inequality (12) is true, then:

$$\text{a) } \bar{N}(0, 1) = \frac{H_2 + H_3}{H_5 + H_6} > \bar{N}(0, 0) = \frac{H_3}{H_6} \tag{13}$$

$$\text{b) } \bar{N}(1, 1) = \frac{H_1 + H_2 + H_3}{H_4 + H_5 + H_6} > \bar{N}(0, 1) = \frac{H_2 + H_3}{H_5 + H_6}$$

$$\text{c) } \bar{N}(1, 1) = \frac{H_1 + H_2 + H_3}{H_4 + H_5 + H_6} > \bar{N}(1, 0) = \frac{H_1 + H_3}{H_4 + H_6}$$

These inequalities are equivalent to:

$$\text{a) } H_2 H_6 > H_3 H_5 \tag{14}$$

$$\text{b) } H_1 (H_5 + H_6) > H_4 (H_2 + H_3)$$

$$\text{c) } H_2 (H_4 + H_6) > H_5 (H_1 + H_3)$$

If $H_1 H_6 > H_3 H_4$, then from eqn (3a), it is clear that for an ergodic system it must be true that:

$$F_M(v_1, v_2) > 0 \tag{15}$$

and, hence, by substitution of this result into eqn (3b), that $H_2 H_6 - H_3 H_5 > 0$. This proves inequality (14a). Inequalities (14b) and (14c) may be rewritten:

$$\text{a) } H_1 H_6 - H_3 H_4 > -(H_1 H_5 - H_2 H_4) \tag{16}$$

$$\text{b) } H_2 H_6 - H_3 H_5 > H_1 H_5 - H_2 H_4$$

These inequalities were verified in the proof of theorem 4.1-1 in the form:

$$\text{a) } H_1 H_6 - H_3 H_4 > |H_1 H_5 - H_2 H_4| \tag{17}$$

$$\text{b) } H_2 H_6 - H_3 H_5 > |H_1 H_5 - H_2 H_4|$$

This completes the proof of case 1.

Case 2 is completely analogous to case 1 with the ">" sign replaced by "<". It leads to the same inequalities and, hence, is proven as a consequence of case 1. This completes the proof of Theorem 4.1-1.

Having shown that the probabilistic model is sub-optimal, the equation for $\bar{N}$ using a deterministic model with $\alpha = \beta = 0$ becomes:

$$\bar{N}_M(v_1, v_2) = \frac{(v_2 + v_1 - 1)^2 \left[ v_2 \sum_{i=1}^{M} \sum_{j=i}^{M} i v_1^{j-i} f_j - \sum_{i=0}^{M-1} f_i + (M+1) f_M \right] + [(M+2)(v_2 + v_1 - 1) + 1](f_M - f_{M-1})}{(v_2 + v_1 - 1) \left[ \left( v_2 \sum_{i=0}^{M} \sum_{j=0}^{i} v_1^j f_i + f_M \right) (v_2 + v_1 - 1) + f_M - f_{M-1} \right]} \tag{18}$$

To understand the nature of $\bar{N}_M(v_1, v_2)$, we will examine how it changes with M, showing that it can be written in the form:

$$\bar{N}_M(v_1, v_2) = \frac{h_1(v_1, v_2) + \sum_{i=0}^{M} \Delta_{num}(i, v_1, v_2)}{h_2(v_1, v_2) + \sum_{i=0}^{M} \Delta_{denom}(i, v_1, v_2)} \tag{19}$$

We will consider first the numerator and begin by taking the difference between the numerators for $M = \hat{M}$ and $M = \hat{M} - 1$:

$$\Delta_{num}(\hat{M}, v_1, v_2) = (v_2 + v_1 - 1)^2 \left[ v_2 \sum_{i=1}^{\hat{M}} i v_1^{\hat{M}-i} f_{\hat{M}} - f_{\hat{M}-1} + (\hat{M} + 1) f_{\hat{M}} - \hat{M} f_{\hat{M}-1} \right] \tag{20}$$

$$+ [(\hat{M} + 2)(v_2 + v_1 - 1) + 1] (f_{\hat{M}} - f_{\hat{M}-1}) - [(\hat{M} + 1)(v_2 + v_1 - 1) + 1] (f_{\hat{M}-1} - f_{\hat{M}-2})$$

$$= (v_2 + v_1 - 1)^2 \left[ v_2 \sum_{i=1}^{\hat{M}} i v_1^{\hat{M}-i} f_{\hat{M}} + (\hat{M} + 1)(f_{\hat{M}} - f_{\hat{M}-1}) \right]$$

$$+ [(\hat{M} + 1)(v_2 + v_1 - 1) + 1] (f_{\hat{M}} - 2f_{\hat{M}-1} + f_{\hat{M}-2}) + (f_{\hat{M}} - f_{\hat{M}-1})(v_2 + v_1 - 1)$$

Using the identity given as eqn (10) in Appendix B and repeated here:

$$v_1 f_i = (v_2 + v_1 + 1) f_{i-1} - f_{i-2} \tag{21}$$

eqn (20) can be simplified to:

$$\Delta_{num}(\hat{M}, v_1, v_2) = \left[ (v_2 + v_1 - 1)^2 \sum_{i=1}^{\hat{M}} i v_1^{\hat{M}-i} + (\hat{M} + 1)(v_2 + v_1 - 1) + 1 \right] v_2 f_{\hat{M}} \tag{22}$$

A similar process for the denominator yields:

$$\Delta_{denom}(\hat{M}, v_1, v_2) = (v_2 + v_1 - 1) \left[ (v_2 + v_1 - 1) \sum_{j=0}^{\hat{M}} v_1^j + 1 \right] v_2 f_{\hat{M}} \tag{23}$$

Eqns (22) and (23) used in eqn (18) yield the desired form:

$$\bar{N}_M(v_1, v_2) = \frac{v_1(v_2 + v_1) + \sum_{i=0}^{M} \left[ (v_2 + v_1 - 1)^2 \sum_{j=0}^{i} j v_1^{i-j} + (i + 1)(v_2 + v_1 - 1) + 1 \right] v_2 f_i}{(v_2 + v_1 - 1) \left\{ v_1 + \sum_{i=0}^{M} \left[ (v_2 + v_1 - 1) \sum_{j=0}^{i} v_1^j + 1 \right] v_2 f_i \right\}} \tag{24}$$

This expression for $\bar{N}$ clearly expresses its dependence on the threshold size M, which will be treated as the control parameter for the optimization process to be considered next.

4.2 Finding the Optimal Threshold Size

It was shown in section 4.1 that the mean number of customers in the system ($\bar{N}$) is given by the following equation, in which the variable M is the threshold

$$\bar{N}_M(v_1, v_2) = \frac{v_1(v_2 + v_1) + \sum_{i=0}^{M} \left[ (v_2 + v_1 - 1)^2 \sum_{j=0}^{i} j v_1^{i-j} + (i+1)(v_2 + v_1 - 1) + 1 \right] v_2 f_i}{(v_2 + v_1 - 1) \left\{ v_1 + \sum_{i=0}^{M} \left[ (v_2 + v_1 - 1) \sum_{j=0}^{i} v_1^j + 1 \right] v_2 f_i \right\}} \quad (1)$$

size and is the control variable which facilitates the minimization of $\bar{N}$ for a given $v_1$ and $v_2$. In this section, minimization of $\bar{N}$ is explored, beginning with the following results which explicate the behavior of eqn (1), including a mathematical interpretation of the M = 0 case which gives preemptive preference to server 2. Appendix E shows the results also hold for a non-preemptive interpretation of the M = 0 case.

Lemma 4.2-1 A threshold value of M = 1 results in a lower $\bar{N}$ than M = 0 for an ergodic system with $v_1 > v_2$.

Proof

a) $$\bar{N}_0(v_1, v_2) = \frac{(v_2 + v_1)^2}{(v_2 + v_1 - 1)\,[v_1 + v_2(v_2 + v_1)]} \quad (2)$$

b) $$\bar{N}_1(v_1, v_2) = \frac{(v_2 + v_1)^2 \left[ 1 + \frac{v_2}{v_1} (v_2 + v_1 + 1) \right]}{(v_2 + v_1 - 1) \left\{ v_1 + v_2(v_2 + v_1) + \frac{v_2}{v_1} [(v_2 + v_1 - 1)(v_1 + 1) + 1] (v_2 + v_1 + 1) \right\}}$$

We note from eqn (1) that $\bar{N}_M(v_1, v_2)$ is of the form:

$$\bar{N}_M(v_1, v_2) = \frac{h_1(v_1, v_2) + \sum_{i=0}^{M} \Delta_i^n (v_1, v_2)}{h_2(v_1, v_2) + \sum_{i=0}^{M} \Delta_i^d (v_1, v_2)} \quad (3)$$

where both $\Delta_i^n(v_1, v_2)$ and $\Delta_i^d(v_1, v_2)$ are strictly positive functions

and

$$\text{a) } \Delta^n_{i+1}(v_1, v_2) > \Delta^n_i(v_1, v_2) \tag{4}$$

$$\text{b) } \Delta^d_{i+1}(v_1, v_2) > \Delta^d_i(v_1, v_2)$$

Referring to eqn (2), we have:

$$\text{a) } \Delta^n_1(v_1, v_2) = \frac{v_2}{v_1}\,(v_2 + v_1)^2\,(v_2 + v_1 + 1) \tag{5}$$

$$\text{b) } \Delta^d_1(v_1, v_2) = \frac{v_2}{v_1}\,(v_2 + v_1 - 1)\,(v_2 + v_1 + 1)\,[(v_2 + v_1 - 1)\,(v_1 + 1) + 1]$$

Now for $\bar{N}_1(v_1, v_2)$ to be less than $\bar{N}_0(v_1, v_2)$, the following condition must be satisfied:

$$\frac{\Delta^n_1(v_1, v_2)}{\Delta^d_1(v_1, v_2)} < \bar{N}_0(v_1, v_2) \tag{6}$$

or

$$\frac{(v_2 + v_1)^2}{(v_2 + v_1 - 1)\,[(v_2 + v_1 - 1)\,(v_1 + 1) + 1]} = \frac{(v_2 + v_1)^2}{(v_2 + v_1 - 1)\,[v_2 + v_1(v_2 + v_1)]} \tag{7}$$

$$< \frac{(v_2 + v_1)^2}{(v_2 + v_1 - 1)\,[v_1 + v_2(v_2 + v_1)]}$$

Since $v_1 + v_2(v_2 + v_1) > v_2 + v_1(v_2 + v_1)$ by assumption, inequality (7) holds, and, therefore, the lemma is true. The meaning of this lemma is simply that a customer arriving to an empty system should be sent to the faster of the two servers.

The generalization of condition (6) can be stated as follows. If $\bar{N}_{M+1}(v_1, v_2) \left\{ \begin{matrix} < \\ = \\ > \end{matrix} \right\} \bar{N}_M(v_1, v_2)$, then $\dfrac{\Delta^n_{M+1}(v_1, v_2)}{\Delta^d_{M+1}(v_1, v_2)} \left\{ \begin{matrix} < \\ = \\ > \end{matrix} \right\} \bar{N}_M(v_1, v_2)$. The ratio of $\Delta^n_M(v_1, v_2)$ to $\Delta^d_M(v_1, v_2)$ is clearly an important quantity with respect to the behavior of $\bar{N}$. This ratio is explored further in the following lemma.

Lemma 4.2-2 $\dfrac{\Delta^n_{M+1}(v_1, v_2)}{\Delta^d_{M+1}(v_1, v_2)} > \dfrac{\Delta^n_M(v_1, v_2)}{\Delta^d_M(v_1, v_2)}$ for all $M \geq 0$ in an ergodic system with $v_1 > v_2$.

Proof Substituting for the $\Delta_M^n$ and $\Delta_M^d$, the lemma can be restated as:

$$\frac{(v_2 + v_1 - 1)^2 \sum_{j=0}^{M+1} j v_1^{M+1-j} + (M+2)(v_2 + v_1 - 1) + 1}{(v_2 + v_1 - 1)^2 \sum_{j=0}^{M+1} v_1^j + v_2 + v_1 - 1} > \frac{(v_2 + v_1 - 1)^2 \sum_{j=0}^{M} j v_1^{M-j} + (M+1)(v_2 + v_1 - 1) + 1}{(v_2 + v_1 - 1)^2 \sum_{j=0}^{M} v_1^j + v_2 + v_1 - 1} \quad (8)$$

or

$$(v_2 + v_1 - 1)^2 \sum_{j=0}^{M+1} \sum_{k=0}^{M} j v_1^{M+1+k-j} + (v_2 + v_1 - 1)\left[\sum_{j=0}^{M+1} j v_1^{M+1-j} + (M+2)\sum_{j=0}^{M} v_1^j\right] + M + 2 + \sum_{j=0}^{M} v_1^j$$

$$> (v_2 + v_1 - 1)^2 \sum_{j=0}^{M} \sum_{k=0}^{M+1} j v_1^{M+k-j} + (v_2 + v_1 - 1)\left[\sum_{j=0}^{M} j v_1^{M-j} + (M+1)\sum_{j=0}^{M+1} v_1^j\right] + M + 1 + \sum_{j=0}^{M+1} v_1^j \quad (9)$$

which simplifies to the condition:

$$(v_2 + v_1 - 1)^2 \sum_{j=0}^{M} (j+1) v_1^j + 2(v_2 + v_1 - 1) \sum_{j=0}^{M} v_1^j + 1 > [(M+1)(v_2 + v_1 - 1) + 1] v_1^{M+1} \quad (10)$$

The proof of inequality (10) proceeds by induction on M. We begin with the M = 0 case:

$$(v_2 + v_1 - 1)^2 + 2(v_2 + v_1 - 1) + 1 = (v_2 + v_1)^2 > (v_2 + v_1) v_1 \quad (11)$$

or

$$v_2 + v_1 > v_1$$

which is clearly true.

For the inductive step, assume inequality (10) holds for M = m:

$$(v_2 + v_1 - 1)^2 \sum_{j=0}^{m} (j+1) v_1^j + 2(v_2 + v_1 - 1) \sum_{j=0}^{m} v_1^j + 1 - [(m+1)(v_2 + v_1 - 1) + 1] v_1^{m+1} > 0 \quad (12)$$

For M = m + 1, we have:

$$(v_2 + v_1 - 1)^2 \sum_{j=0}^{m+1} (j+1) v_1^j + 2(v_2 + v_1 - 1) \sum_{j=0}^{m+1} v_1^j + 1 - [(m+2)(v_2 + v_1 - 1) + 1] v_1^{m+2} > 0 \quad (13)$$

Constructing the difference between inequalities (13) and (12) and simplifying yields the following inequality:

$$[(m+2)(v_2+v_1-1)+1]\, v_1^{m+1} v_2 > 0 \tag{14}$$

which is true, completing the proof of lemma 4.2-2. The implication of this lemma is that if $\bar{N}_{M+1}(v_1, v_2) > \bar{N}_M(v_1, v_2)$, then $\bar{N}_{M+1+i}(v_1, v_2) > \bar{N}_{M+i}(v_1, v_2)$ for all $i \geq 0$. These two lemmas show that $\bar{N}_i(v_1, v_2)$, $i = 0, 1, 2, \ldots$, decreases initially with respect to i, and that if it ever begins to increase as i increases, then it continues to increase with i forever. These two lemmas from the basis for proving the uniqueness of an optimal threshold. Existence of an optimal threshold follows from the next lemma:

Lemma 4.2-3 There exists a finite M for which $\bar{N}_{M+1}(v_1, v_2) > \bar{N}_M(v_1, v_2)$.

Proof The proof proceeds by showing that there exists an M such that:

$$\frac{\Delta^n_{M+1}(v_1, v_2)}{\Delta^d_{M+1}(v_1, v_2)} > \bar{N}_M(v_1, v_2) \tag{15}$$

Substituting appropriately into inequality (15) yields:

$$\frac{(v_2+v_1-1)^2 \sum_{j=0}^{M+1} j v_1^{M+1-j} + (M+2)(v_2+v_1-1)+1}{(v_2+v_1-1)^2 \sum_{j=0}^{M+1} v_1^j + v_2 + v_1 - 1}$$

$$> \frac{v_1(v_2+v_1) + \sum_{i=0}^{M} \left[ (v_2+v_1-1)^2 \sum_{j=0}^{i} j v_1^{i-j} + (i+1)(v_2+v_1-1)+1 \right] v_2 f_i}{(v_2+v_1-1)\left\{ v_1 + \sum_{i=0}^{M} \left[ (v_2+v_1-1) \sum_{j=0}^{i} v_1^j + 1 \right] v_2 f_i \right\}} \tag{16}$$

Grouping terms and simplifying inequality (16) results in:

$$\sum_{i=0}^{M} \Bigg\{ (v_2+v_1-1)^2 v_2 \sum_{j=0}^{M+1} \sum_{k=0}^{i} (j v_1^{M+1-j+k} - k v_1^{j+i-k}) + (v_2+v_1-1) v_2 \sum_{j=0}^{M+1} [j v_1^{M+1-j} - (i+1) v_1^j]$$

$$+ (v_2+v_1-1) v_2 \sum_{j=0}^{i} (M+2-j) v_1^{i-j} + (M+1-i) v_2 - v_2 \sum_{j=i+1}^{M+1} v_1^j \Bigg\} f_i$$

$$+ (v_2+v_1-1) v_1 \sum_{i=0}^{M+1} (i-1) v_1^{M+1-i} + (M+1) v_1 - v_1 \sum_{i=0}^{M+1} v_1^i > 0 \tag{17}$$

which can be further simplified to yield:

$$\left(\frac{v_2}{v_1}\right)^2 \sum_{i=0}^{M} \left\{ v_2 \sum_{j=0}^{M-i} (i+1)(j+1) v_1^{-j} + v_2 \sum_{j=0}^{i-1} (M+1-i)(j+1) v_1^{j-M} + v_1 \sum_{j=0}^{M-i} (i+1+j) v_1^{-j} \right\} f_i$$

$$+ \frac{v_2}{v_1} \sum_{i=0}^{M} i v_1^{-i} > v_2 + v_1 \text{ for some M.} \tag{18}$$

Recalling that $f_i \geq 1$ for all $i \geq 0$ and dropping terms of the form $v_1^{-i}$ for $i \geq 1$, we may write:

$$\left(\frac{v_2}{v_1}\right)^2 \sum_{i=0}^{M} \left\{ v_2 \sum_{j=0}^{M-i} (i+1)(j+1) v_1^{-j} + v_2 \sum_{j=0}^{i-1} (M+1-i)(j+1) v_1^{j-M} + v_1 \sum_{j=0}^{M-i} (i+1+j) v_1^{-j} \right\} f_i$$

$$+ \frac{v_2}{v_1} \sum_{i=0}^{M} i v_1^{-i}$$

$$> \left(\frac{v_2}{v_1}\right)^2 \sum_{i=0}^{M} [v_2(i+1) + v_1(i+1)]$$

$$= \left(\frac{v_2}{v_1}\right)^2 (v_2 + v_1) \sum_{i=1}^{M+1} i > v_2 + v_1 \text{ for some M.} \tag{19}$$

It is quite apparent at this point that if:

$$\sum_{i=1}^{M+1} i > \left(\frac{v_1}{v_2}\right)^2 \tag{20}$$

for some M, then inequality (18) is true for that same M, and the lemma is proven. Clearly, given values for $v_1$ and $v_2$, one can select an M to make (20) hold. This completes the proof of lemma 4.2–3.

Inequality (20) can be further exploited to establish an upper bound on the optimal M. Denoting this upper bound by $\hat{M}$,

$$\hat{M} = \left\{ \text{minimum M} \;\middle|\; \sum_{i=1}^{M+1} i > \left(\frac{v_1}{v_2}\right)^2 \right\} \tag{21}$$

Since

$$\sum_{i=1}^{M+1} i = \frac{1}{2}(M + 1)(M + 2) \tag{22}$$

we have:

$$\hat{M} = \left\{ \text{minimum } M \;\middle|\; (M + 1)(M + 2) > 2\left(\frac{v_1}{v_2}\right)^2 \right\} \tag{23}$$

The inequality in (23) can be solved directly to yield the desired bound:

$$\hat{M} = \left\lceil \frac{1}{2}\left(-3 + \sqrt{1 + 8\left(\frac{v_1}{v_2}\right)^2}\right) \right\rceil \leq \left\lceil \sqrt{2}\,\frac{v_1}{v_2} \right\rceil \tag{24}$$

Lemmas 4.2-1, 4.2-2, and 4.2-3 provide the foundation for the following theorem.

Theorem 4.2-1 There exists an $M^* > 0$ such that:

a) $\bar{N}_M(v_1, v_2) > \bar{N}_{M^*}(v_1, v_2)$ for $0 \leq M < M^*$ (25)

b) $\bar{N}_{M^*}(v_1, v_2) \leq \bar{N}_{M^*+1}(v_1, v_2)$

and c) $\bar{N}_{M^*}(v_1, v_2) < \bar{N}_M(v_1, v_2)$ for $M > M^* + 1$

Proof Lemmas 4.2-1, 4.2-2, and 4.2-3 respectively show that $\bar{N}_M(v_1, v_2)$, $M = 0, 1, \ldots$ starts out as a decreasing sequence, that if it ever begins to increase then it remains a monotonically increasing sequence beyond that point, and that there exists a point ($M^* + 1$ or $M^* + 2$) at which the sequence begins to increase. The possibility exists that

$$\frac{\Delta^n_{M^*+1}(v_1, v_2)}{\Delta^d_{M^*+1}(v_1, v_2)} = \bar{N}_{M^*}(v_1, v_2) \tag{26}$$

which results in $\bar{N}_M(v_1, v_2)$ taking on its minimum at two consecutive points ($M = M^*$ and $M = M^* + 1$). Lemma 4.2-2 proves that in this event.

$$\bar{N}_{M^*+2}(v_1, v_2) > \bar{N}_{M^*+1}(v_1, v_2) \tag{27}$$

so we see that the minimum $\bar{N}_M(v_1, v_2)$ occurs at either a unique M or at two consecutive values of M. This completes the proof of the theorem.

Having proven the existence and quasi-uniqueness of a minimum $\bar{N}_M(v_1, v_2)$, we now wish to find the M = M* which produces this minimum. A straightforward linear search strategy will work, as shown in the following algorithm:

Algorithm 4.2-1 Search for optimal M

[1] $M \leftarrow 1$

[2] While $\bar{N}_{M+1}(v_1, v_2) < \bar{N}_M(v_1, v_2)$

Do Begin

$M \leftarrow M + 1$

End

[3] $M^* \leftarrow M$

[4] Stop; M* minimizes $\bar{N}_M(v_1, v_2)$.

Since we have established an upper bound $\hat{M}$ on M*, algorithm 4.2-1 is guaranteed to terminate in finite time. The computation time can be reduced, however, by utilizing eqn (18). The optimal M is the minimum M which satisfies the inequality:

$$\left(\frac{v_2}{v_1}\right)^2 \sum_{i=0}^{M} \left\{ v_2 \sum_{j=0}^{M-i} (i+1)(j+1) v_1^{-j} + v_2 \sum_{j=0}^{i-1} (M+1-i)(j+1) v_1^{j-M} + v_1 \sum_{j=0}^{M-i} (i+1+j) v_1^{-j} \right\} f_i$$

$$+ \frac{v_2}{v_1} \sum_{i=0}^{M} i v_1^{-i} \geq v_2 + v_1 \qquad (28)$$

Algorithm 4.2-2 is derived from (28) and runs substantially faster than Algorithm 4.2-1.

Algorithm 4.2-2 Faster Search for Optimal Threshold M*

[1] $M \leftarrow 0$; SUMTERMS $\leftarrow \left(\frac{v_2}{v_1}\right)^2 (v_2 + v_1)$

[2] While SUMTERMS $< v_2 + v_1$

Do Begin

$M \leftarrow M + 1$;

$x_1 \leftarrow \left(\frac{v_2}{v_1}\right)^2 v_2 \sum_{i=0}^{M} (i+1)(M+1-i) v_1^{i-M} f_i$;

$$x_2 \leftarrow \left(\frac{v_2}{v_1}\right)^2 v_2 \sum_{i=1}^{M} \sum_{j=0}^{i-1} (M + 1 - i)(j + 1)\, v_1^{j-M} f_i;$$

$$x_3 \leftarrow \left(\frac{v_2}{v_1}\right)^2 v_1 \sum_{i=0}^{M} (M + 1)\, v_1^{i-M} f_i;$$

$$x_4 \leftarrow M\, v_2\, v_1^{-M-1};$$

$$x_5 \leftarrow \left(\frac{v_2}{v_1}\right)^2 v_2 \sum_{i=1}^{M-1} \sum_{j=0}^{i-1} (M - i)(j + 1)\, v_1^{j-M+1} f_i;$$

$$\text{SUMTERMS} \leftarrow \text{SUMTERMS} + \sum_{i=1}^{4} x_i - x_5;$$

End

[3] $M^* \leftarrow M$;

[4] Stop; $M^*$ minimizes $\bar{N}_M(v_1, v_2)$

Figure 4.2–1 displays the optimal threshold $M^*$ as a function of $v_1$ and $v_2$. We note that this takes the form of a partitioning of the $(v_1, v_2)$ plane into semi-infinite regions of constant $M^*$, and that the boundaries between the $M^*$ and $M^* + 1$ regions appear to approach asymptotes of slope $\frac{1}{M^* + 1}$. That this is, in fact, the case is shown in the following lemma.

Lemma 4.2–4 The slope of the boundary between the region of the $(v_1, v_2)$ plane where the optimal threshold is $M^*$ and the region where the optimal threshold is $M^* + 1$ approaches $\frac{1}{M^* + 1}$ as $v_1 \to \infty$.

Proof The equation for the boundaries comes from eqn (28):

$$\left(\frac{v_2}{v_1}\right)^2 \sum_{i=0}^{M^*} \left\{ \frac{v_2}{v_1} \sum_{j=0}^{M^*-i} (i+1)(j+1) v_1^{-j} + \frac{v_2}{v_1} \sum_{j=0}^{i-1} (M^*+1-i)(j+1) v_1^{j-M^*} + \sum_{j=0}^{M^*-i} (i+1+j) v_1^{-j} \right\} f_i$$

$$+ \frac{v_2}{v_1} \sum_{i=0}^{M^*} i v_1^{-i-1} - \frac{v_2}{v_1} - 1 = 0 \qquad (29)$$

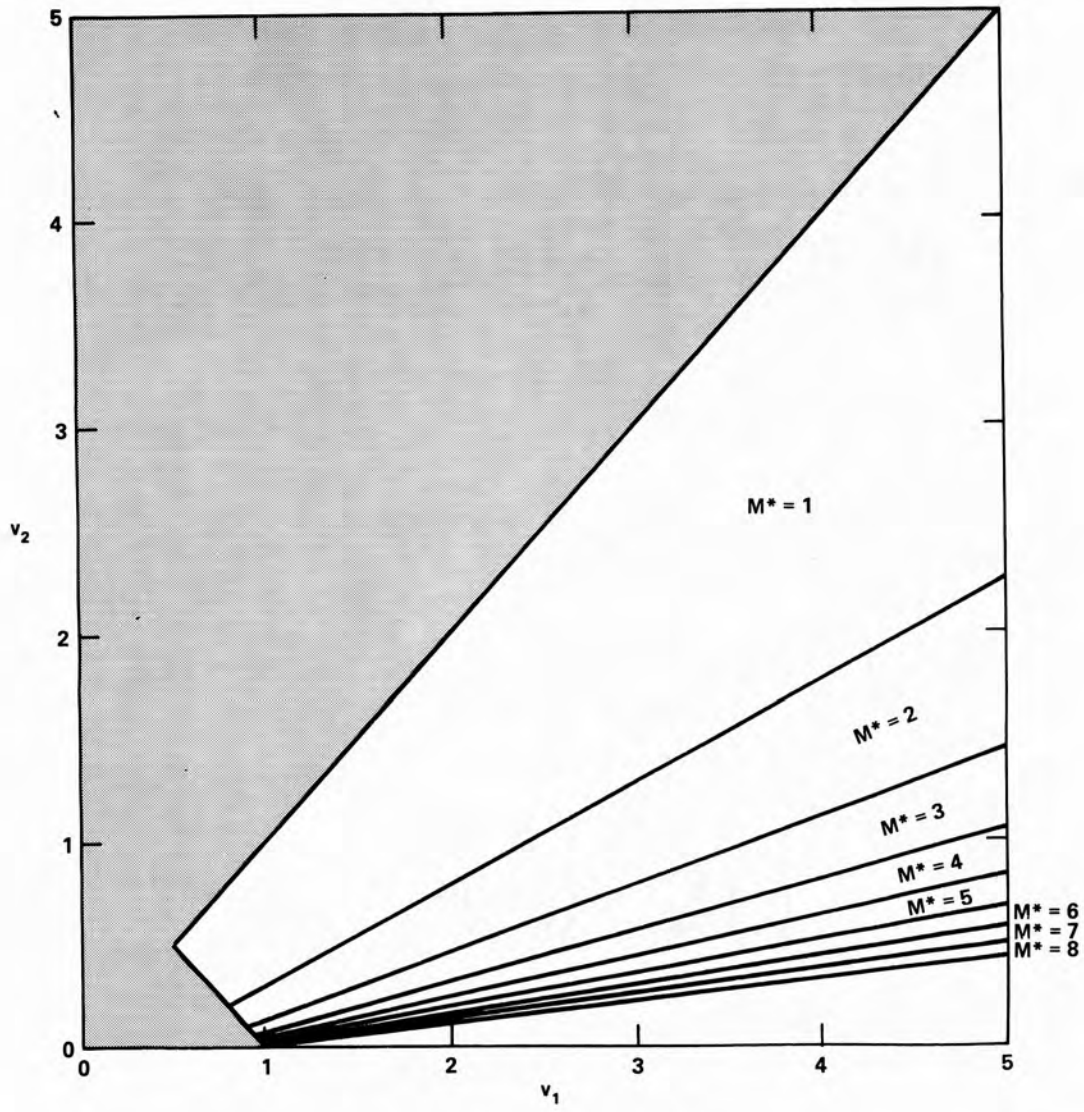



Figure 4.2-1. Optimal Threshold Partitions

Now for $f_i$, we have:

$$\lim_{v_1 \to \infty} f_i = \lim_{v_1 \to \infty} f_i \left( \frac{v_2}{v_1} + 1 + \frac{1}{v_1}, \frac{1}{v_1} \right)$$

$$= \lim_{v_1 \to \infty} \sum_{j=0}^{\lfloor \frac{i}{2} \rfloor} (-1)^j \binom{i-j}{j} \left( \frac{v_2}{v_1} + 1 + \frac{1}{v_1} \right)^{i-2j} \left( \frac{1}{v_1} \right)^j$$

$$= \left( \frac{v_2}{v_1} + 1 \right)^i \tag{30}$$

Taking the limit of eqn (29):

$$\lim_{v_1 \to \infty} \left\{ \left( \frac{v_2}{v_1} \right)^2 \sum_{i=0}^{M^*} \left[ \frac{v_2}{v_1} \sum_{j=0}^{M^*-i} (i+1)(j+1) v_1^{-j} + \frac{v_2}{v_1} \sum_{j=0}^{i-1} (M^*+1-i)(j+1) v_1^{j-M^*} + \sum_{j=0}^{M^*-i} (i+1+j) v_1^{-j} \right] f_i \right.$$

$$\left. + \frac{v_2}{v_1} \sum_{i=0}^{M^*} i v_1^{-i-1} - \frac{v_2}{v_1} - 1 \right\}$$

$$= \left( \frac{v_2}{v_1} \right)^2 \sum_{i=0}^{M^*} \left[ \frac{v_2}{v_1} (i+1) + (i+1) \right] \left( \frac{v_2}{v_1} + 1 \right)^i - \left( \frac{v_2}{v_1} + 1 \right)$$

$$= \left( \frac{v_2}{v_1} + 1 \right) \left[ \left( \frac{v_2}{v_1} \right)^2 \sum_{i=0}^{M^*} (i+1) \left( \frac{v_2}{v_1} + 1 \right)^i - 1 \right] = 0 \tag{31}$$

Since $\frac{v_2}{v_1} + 1$ is strictly positive, we have:

$$\left( \frac{v_2}{v_1} \right)^2 \sum_{i=0}^{M^*} (i+1) \left( \frac{v_2}{v_1} + 1 \right)^i - 1 = 0 \tag{32}$$

Letting $x \equiv \frac{v_2}{v_1} + 1$ in eqn (32):

$$(x-1)^2 \sum_{i=0}^{M^*} (i+1) x^i - 1 = (x^2 - 2x + 1) \sum_{i=0}^{M^*} (i+1) x^i - 1$$

$$= \sum_{i=2}^{M^*+2} (i-1) x^i - \sum_{i=1}^{M^*+1} 2i x^i + \sum_{i=0}^{M^*} (i+1) x^i - 1$$

$$= x^{M^*+1} [(M^*+1) x - (M^*+2)] = 0 \tag{33}$$

The desired solution to this equation is:

$$x = \frac{M^* + 2}{M^* + 1} \tag{34}$$

and so

$$\lim_{v_1 \to \infty} \frac{v_2}{v_1} = \frac{M^* + 2}{M^* + 1} - 1 = \frac{1}{M^* + 1} \tag{35}$$

which proves the lemma.

Noting from figure 4.2-1 that the boundaries between regions of constant threshold size appear to be close to linear, we now investigate the utility of the asymptotes to the boundaries as approximations of the boundaries themselves. We begin by deriving the equations defining the asymptotes.

Theorem 4.2-2 The asymptote to the boundary between the region where the optimal threshold size is $M^*$ and the region where the optimal threshold size is $M^* + 1$ is given by the equation:

$$v_2 = \frac{v_1}{M^* + 1} - \frac{M^*}{(M^* + 1)^2} \tag{36}$$

Proof Eqn (29) defines the boundaries and is repeated here:

$$\left(\frac{v_2}{v_1}\right)^2 \sum_{i=0}^{M^*} \left\{ \frac{v_2}{v_1} \sum_{j=0}^{M^*-i} (i+1)(j+1)v_1^{-j} + \frac{v_2}{v_1} \sum_{j=0}^{i-1} (M^*+1-i)(j+1)v_1^{j-M^*} + \sum_{j=0}^{M^*-i} (i+1+j)v_1^{-j} \right\} f_i$$

$$+ \frac{v_2}{v_1} \sum_{i=0}^{M^*} i v_1^{-i-1} - \left(\frac{v_2}{v_1} + 1\right) = 0 \tag{37}$$

The initial approach will be identical to the approach used in the proof of lemma 4.2-4, but this time terms of $o\left(\frac{1}{v_1}\right)$ will be dropped from the limiting form of eqn (37) as $v_1 \to \infty$. For the $f_i$ we have:

a) $f_0 = 1$ (38)

b) $\lim_{v_1 \to \infty} f_i = \left(\frac{v_2}{v_1} + 1\right)^i + \frac{1}{v_1}\left(\frac{iv_2}{v_1} + 1\right)\left(\frac{v_2}{v_1} + 1\right)^{i-2} \quad i \geq 1$

The limiting form of eqn (37) is:

$$\left(\frac{v_2}{v_1}\right)^2 \sum_{i=0}^{M^*} (i+1)\left(\frac{v_2}{v_1}+1\right)^{i+1} - \left(\frac{v_2}{v_1}+1\right) + \frac{1}{v_1}\left(\frac{v_2}{v_1}\right)^2 \sum_{i=0}^{M^*-1} \left[i^2 \frac{v_2}{v_1} + 1 + 2(i+1)\left(\frac{v_2}{v_1}+1\right)^2\right]\left(\frac{v_2}{v_1}+1\right)^{i-1}$$

$$- \frac{1}{v_1}\left(\frac{v_2}{v_1}\right)^2 \left(\frac{v_2}{v_1}+1\right)^{-1} + \frac{1}{v_1}\left(\frac{v_2}{v_1}\right)^2 \left[M^*\left(M^* + \frac{v_2}{v_1} + 1\right)\left(\frac{v_2}{v_1}+1\right) - M^{*2} + 1\right]\left(\frac{v_2}{v_1}+1\right)^{M^*-1} = 0 \quad (39)$$

We will approach the simplification of eqn (39) in two steps. Expanding first the non-$\frac{1}{v_1}$ terms, we get:

$$\left(\frac{v_2}{v_1}\right)^2 \sum_{i=0}^{M^*} (i+1)\left(\frac{v_2}{v_1}+1\right)^{i+1} - \left(\frac{v_2}{v_1}+1\right)$$

$$= \left(\frac{v_2}{v_1}\right)^2 \left[\left(\frac{v_2}{v_1}+1\right) + 2\left(\frac{v_2}{v_1}+1\right)^2 + 3\left(\frac{v_2}{v_1}+1\right)^3 + \ldots + (M^*+1)\left(\frac{v_2}{v_1}+1\right)^{M^*+1}\right] - \left(\frac{v_2}{v_1}+1\right)$$

$$= \left(\frac{v_2}{v_1}+1\right)\left[\left(\frac{v_2}{v_1}\right)^2 - 1 + 2\left(\frac{v_2}{v_1}\right)^2\left(\frac{v_2}{v_1}+1\right) + 3\left(\frac{v_2}{v_1}\right)^2\left(\frac{v_2}{v_1}+1\right)^2 + \ldots + (M^*+1)\left(\frac{v_2}{v_1}\right)^2\left(\frac{v_2}{v_1}+1\right)^{M^*}\right]$$

$$= \left(\frac{v_2}{v_1}+1\right)^2 \left[\left(2\frac{v_2}{v_1} - 1\right)\left(\frac{v_2}{v_1}+1\right) + 3\left(\frac{v_2}{v_1}\right)^2\left(\frac{v_2}{v_1}+1\right) + \ldots + (M^*+1)\left(\frac{v_2}{v_1}\right)^2\left(\frac{v_2}{v_1}+1\right)^{M^*-1}\right]$$

$$= \ldots$$

$$= \left(\frac{v_2}{v_1}+1\right)^{M^*+1} \left[(M^*+1)\left(\frac{v_2}{v_1}\right) - 1\right]\left(\frac{v_2}{v_1}+1\right)$$

$$= \left(\frac{v_2}{v_1}+1\right)^{M^*+2} \left[(M^*+1)\left(\frac{v_2}{v_1}\right) - 1\right] \quad (40)$$

Now, expansion of the $\frac{1}{v_1}$ terms yields:

$$\frac{1}{v_1}\left(\frac{v_2}{v_1}\right)\sum_{i=0}^{2M^*-1}\left[i^2\frac{v_2}{v_1}+1+2(i+1)\left(\frac{v_2}{v_1}+1\right)^2\right]\left(\frac{v_2}{v_1}+1\right)^{i-1}-\frac{1}{v_1}\left(\frac{v_2}{v_1}\right)^2\left(\frac{v_2}{v_1}+1\right)^{-1}$$

$$+\frac{1}{v_1}\left(\frac{v_2}{v_1}\right)^2\left[M^*\left(M^*+\frac{v_2}{v_1}+1\right)\left(\frac{v_2}{v_1}+1\right)-M^{*2}+1\right]\left(\frac{v_2}{v_1}+1\right)^{M^*-1}$$

$$=\frac{1}{v_1}\left(\frac{v_2}{v_1}\right)^2\left\{\left[1+2\left(\frac{v_2}{v_1}+1\right)^2\right]\left(\frac{v_2}{v_1}+1\right)^{-1}+\left[\frac{v_2}{v_1}+1+4\left(\frac{v_2}{v_1}+1\right)^2\right]\right.$$

$$+\left[4\frac{v_2}{v_1}+1+6\left(\frac{v_2}{v_1}+1\right)^2\right]\left(\frac{v_2}{v_1}+1\right)+\left[9\frac{v_2}{v_1}+1+8\left(\frac{v_2}{v_1}+1\right)^2\right]\left(\frac{v_2}{v_1}+1\right)^2$$

$$+\ldots+\left[(M^*-1)^2\left(\frac{v_2}{v_1}\right)+1+2M^*\left(\frac{v_2}{v_1}+1\right)^2\right]\left(\frac{v_2}{v_1}+1\right)^{M^*-2}-\left(\frac{v_2}{v_1}+1\right)^{-1}$$

$$\left.+\left[M^*\left(M^*+\frac{v_2}{v_1}+1\right)\left(\frac{v_2}{v_1}+1\right)-M^{*2}+1\right]\left(\frac{v_2}{v_1}+1\right)^{M^*-1}\right\}$$

$$=\frac{1}{v_1}\left(\frac{v_2}{v_1}\right)^2\left\{4\left(\frac{v_2}{v_1}+1\right)^2+\left(5+9\frac{v_2}{v_1}\right)\left(\frac{v_2}{v_1}+1\right)^2+\left(7+16\frac{v_2}{v_1}\right)\left(\frac{v_2}{v_1}+1\right)^3+\left(9+25\frac{v_2}{v_1}\right)\left(\frac{v_2}{v_1}+1\right)^4\right.$$

$$+\ldots+\left[2M^*-3+(M^*-1)^2\frac{v_2}{v_1}\right]\left(\frac{v_2}{v_1}+1\right)^{M^*-2}-(M^*-1)^2\left(\frac{v_2}{v_1}+1\right)^{M^*-1}$$

$$\left.+M^*(2+M^*)\left(\frac{v_2}{v_1}+1\right)^{M^*}+M^*\left(\frac{v_2}{v_1}+1\right)^{M^*+1}\right\}$$

$$=\frac{1}{v_1}\left(\frac{v_2}{v_1}\right)^2\left(\frac{v_2}{v_1}+1\right)^2\left\{9\left(\frac{v_2}{v_1}+1\right)+\left(7+16\frac{v_2}{v_1}\right)\left(\frac{v_2}{v_1}+1\right)+\left(9+25\frac{v_2}{v_1}\right)\left(\frac{v_2}{v_1}+1\right)^2+\ldots\right.$$

$$+\left[2M^*-3+(M^*-1)^2\frac{v_2}{v_1}\right]\left(\frac{v_2}{v_1}+1\right)^{M^*-4}-(M^*-1)^2\left(\frac{v_2}{v_1}+1\right)^{M^*-3}$$

$$\left.+M^*(2+M^*)\left(\frac{v_2}{v_1}+1\right)^{M^*-2}+M^*\left(\frac{v_2}{v_1}+1\right)^{M^*-1}\right\}$$

$$=\ldots$$

$$=\frac{1}{v_1}\left(\frac{v_2}{v_1}\right)^2\left(\frac{v_2}{v_1}+1\right)^{M^*-2}\left[M^*(2+M^*)\left(\frac{v_2}{v_1}+1\right)^2+M^*\left(\frac{v_2}{v_1}+1\right)^3\right]$$

$$=\frac{1}{v_1}\left(\frac{v_2}{v_1}\right)^2\left(\frac{v_2}{v_1}+1\right)^{M^*}M^*\left(\frac{v_2}{v_1}+M^*+3\right) \qquad (41)$$

Adding eqns (40) and (41) yields the limiting form of eqn (37), where the limit is taken as $v_1 \to \infty$:

$$\left(\frac{v_2}{v_1}\right)^{M^*+2}\left[(M^*+1)\left(\frac{v_2}{v_1}\right)-1\right]+\frac{1}{v_1}\left(\frac{v_2}{v_1}\right)^2\left(\frac{v_2}{v_1}+1\right)^{M^*}M^*\left(\frac{v_2}{v_1}+M^*+3\right)=0 \qquad (42)$$

or, equivalently:

$$\left(\frac{v_2}{v_1}+1\right)^2\left[(M^*+1)\left(\frac{v_2}{v_1}\right)-1\right]+\frac{1}{v_1}\left(\frac{v_2}{v_1}\right)^2 M^*\left(\frac{v_2}{v_1}+M^*+3\right)=0 \qquad (43)$$

From lemma 4.2–4, it is known that the equations for the threshold boundary asymptotes are of the form:

$$\frac{v_2}{v_1}=\frac{1}{M^*+1}-\frac{k_{M^*}}{v_1} \qquad M^*=0,1,2,\ldots \qquad (44)$$

where the $k_{M^*}$ are constants to be determined. Substituting eqn (44) into eqn (43) produces:

$$\left(M^*+\frac{1}{M^*+1}+3\right)k_{M^*}-\left(1+\frac{1}{M^*+1}\right)\left[1-\left(\frac{1}{M^*+1}\right)^2\right]=o\left(\frac{1}{v_1}\right) \qquad (45)$$

which can be written:

$$(M^*+1)^2(M^*+2)^2k_{M^*}-M^*(M^*+2)^2=o\left(\frac{1}{v_1}\right) \qquad (46)$$

or

$$(M^*+1)^2k_{M^*}-M^*=o\left(\frac{1}{v_1}\right) \qquad (47)$$

Taking the limit as $v_1 \to \infty$, we see that the $k_{M^*}$ are:

$$k_{M^*}=\frac{M^*}{(M^*+1)^2} \qquad (48)$$

Substituting $k_{M^*}$ back into eqn (44) yields the asymptotic threshold boundary equations and completes the proof of theorem 4.2–2.

$$v_2=\frac{v_1}{M^*+1}-\frac{M^*}{(M^*+1)^2} \qquad (49)$$

Using these asymptotes as approximations to the actual boundaries results in the following approximation $\tilde{M}$ to the optimal threshold size M*:

$$M^* \approx \tilde{M} = \left\lfloor \frac{v_1 - 1 + \sqrt{4v_2 + (v_1 - 1)^2}}{2v_2} \right\rfloor \tag{50}$$

In figure 4.2–2, the asymptotes are superimposed on the actual boundaries between the regions of constant M. It is apparent from the diagram that the asymptotes are quite good approximations to the true boundaries, with the greatest divergence as $v_1 + v_2$ approaches 1, i.e., when the system approaches saturation. Figure 4.2–3 shows a blow–up of this saturation region. In section 4.3, the magnitude of the errors induced by using $\tilde{M}$ rather than M* will be explored.

Figure 4.2–1 is also useful to provide insight into the mechanism by which M* changes for a given system as a function of the customer arrival rate. Recalling that $v_1 = \frac{\mu_1}{\lambda}$ and $v_2 = \frac{\mu_2}{\lambda}$, where $\mu_i$ is the service rate of server i and $\lambda$ is the customer arrival rate, we see that the lines given by the equation:

$$\frac{v_2}{v_1} = \frac{\mu_2}{\mu_1} = \text{constant} \tag{51}$$

represent a server configuration, with each point on the line corresponding to a specific value of $\lambda$. These lines are, of course, simply radials emanating from the origin. We note that at the intersection of a radial and the line $v_1 + v_2 = \frac{\mu_1}{\lambda} + \frac{\mu_2}{\lambda} = 1$, or $\mu_1 + \mu_2 = \lambda$, system saturation is encountered. As we move out along the radial, $\lambda$ decreases, approaching 0 as $v_1$ and $v_2$ approach $\infty$. In figure 4.2–4, a few of these radial lines have been superimposed on the boundaries as shown in figure 4.2–1. It is clear from this diagram that M* takes its lowest value as $\lambda \to \mu_1 + \mu_2$, i.e., as the system approaches saturation. As $\lambda$ decreases, we move out along the radial and cross into regions of progressively higher M*, until we enter the limiting M* region given by:

$$M^*_{max} = \left\lfloor \frac{v_1}{v_2} \right\rfloor \tag{52}$$

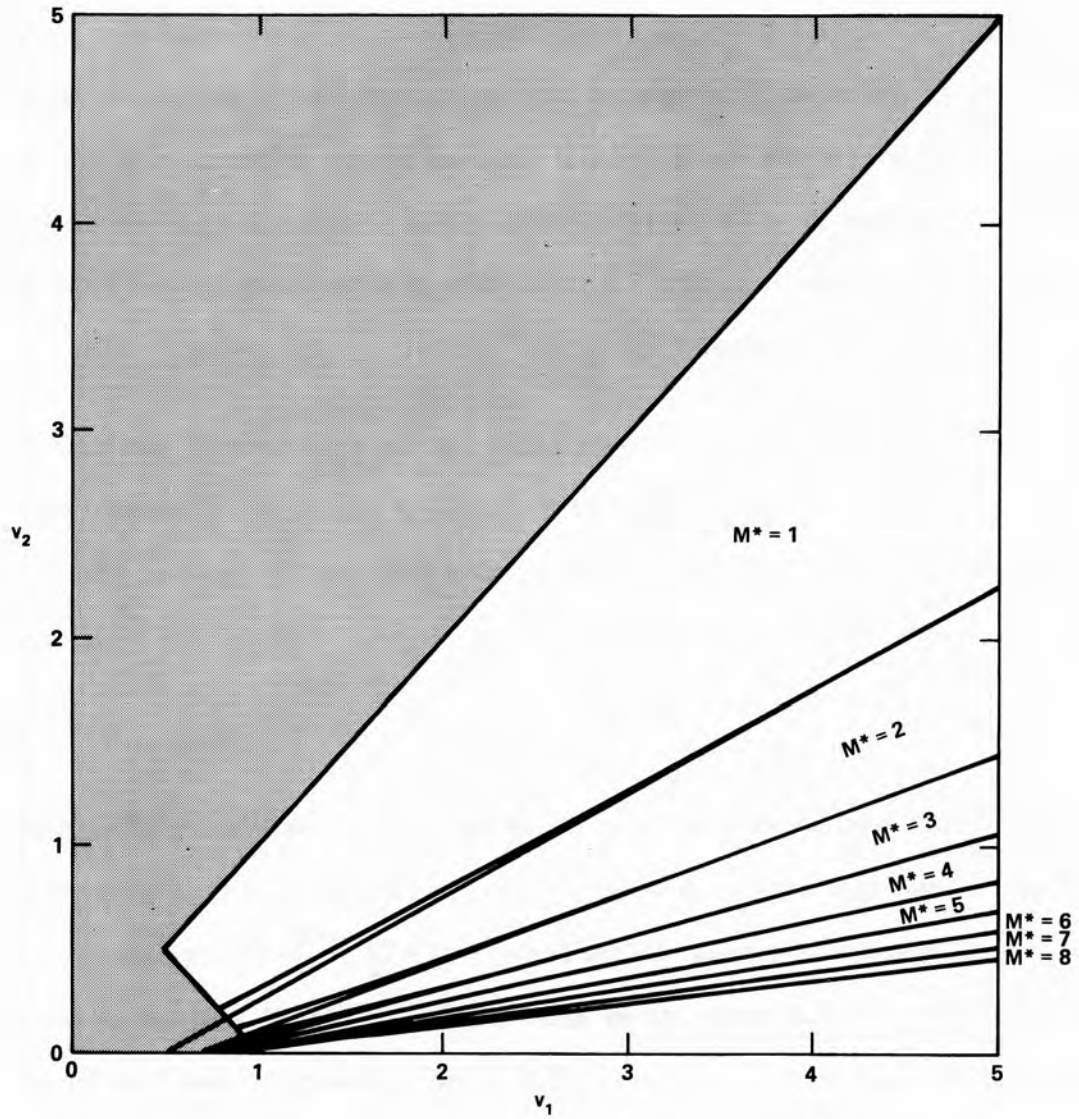



Figure 4.2–2. Optimal Threshold Partitions with Asymptotes

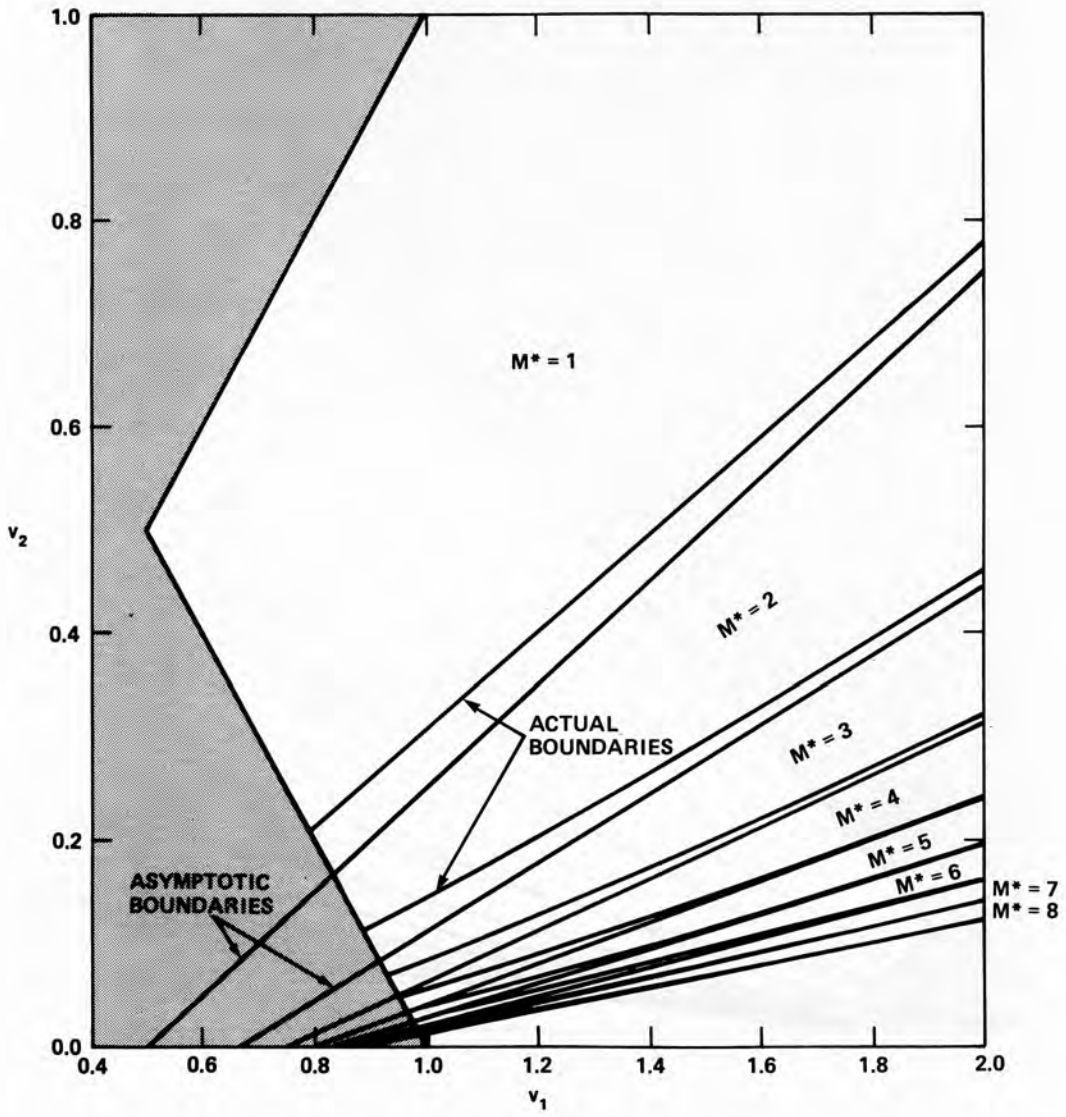



Figure 4.2-3. M* Regions at Saturation

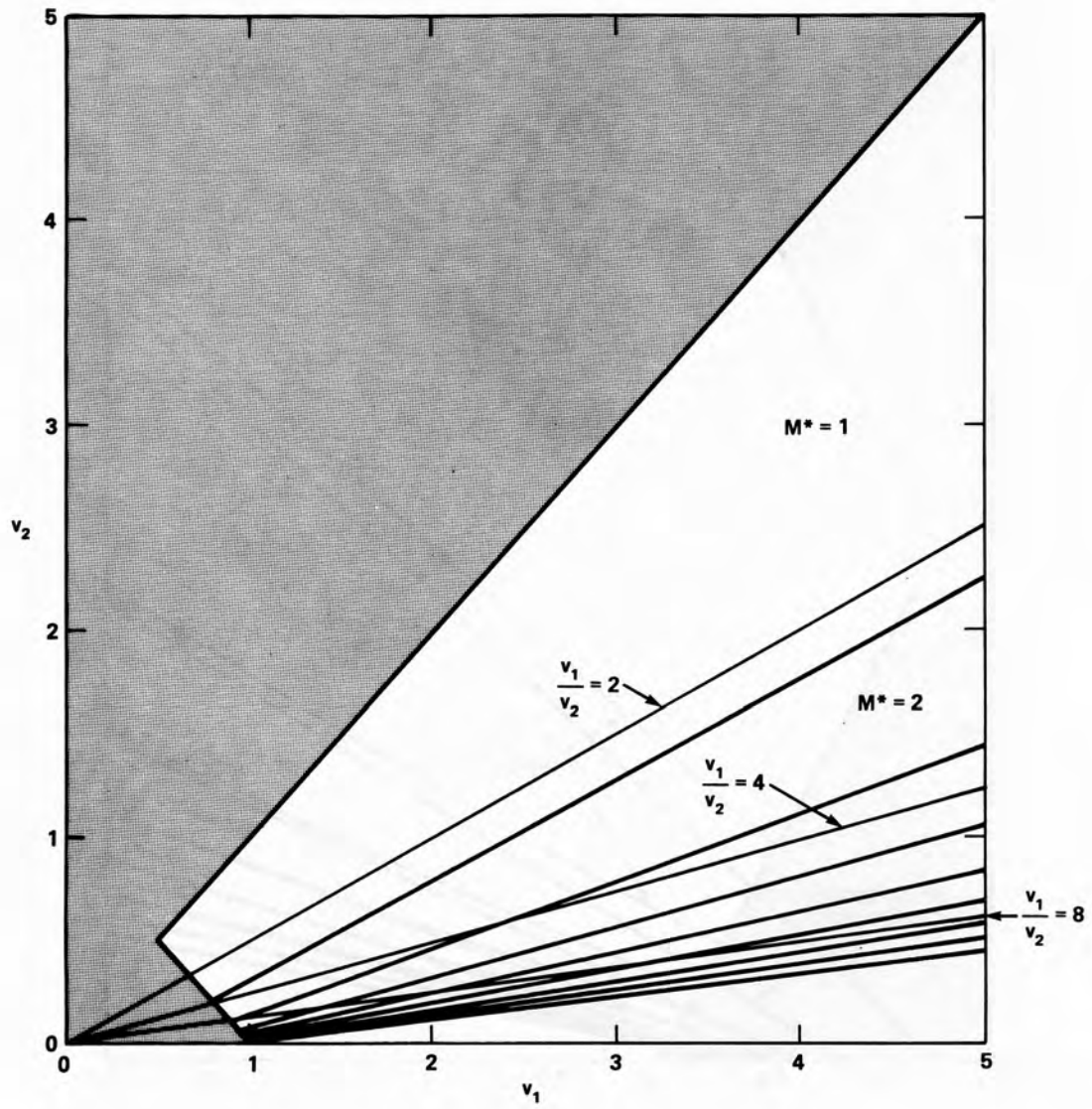



Figure 4.2–4. M* Varies with λ along Radials

## 4.3 Performance Analysis

The initial discussion of the threshold queueing model as presented in section 3.3 introduced four variables to describe the problem: $\mu_1$, $\mu_2$, $\lambda$, and M. Subsequently it was observed that $\lambda$ could be used as a normalization coefficient, and, hence, that the problem could be treated with the three variables $v_1$, $v_2$, and M. In analyzing the performance of the system, it will frequently be more natural to speak in terms of one or the other set of variables. The four variable set has the advantage of correlation to the physical parameters of the system, while the second enables some more general statements to be made about the mathematical nature of the system performance. In the subsequent discussion, both sets of variables will be used, as the situation warrants. As a reminder, the variables carry the following interpretations:

$\mu_i$ = Service rate of server i

$\lambda$ = Customer arrival rate

M = Threshold size

$v_i = \frac{\mu_i}{\lambda}$

The primary performance measure to be employed in this section is the mean customer time in the system ($\overline{T}$). This is directly related to the mean number of customers in the system ($\overline{N}$) by Little's law [LITTJ61]:

$$\overline{T} = \frac{\overline{N}}{\lambda} \qquad (1)$$

Appendices A and B present the derivation of the steady state probabilities associated with the states of the Markov chain representation of the system. These probabilities can be viewed in either of two ways, namely, either the instantaneous probability that the system will be in any given state, or the long term expected proportion of time that the system will be in a given state. Due to the ergodicity of the system both views are equivalent. Figure 4.3-1 (a-j) displays the steady state probabilities for a "typical" system configuration as the threshold value M varies incrementally from one to ten (i.e., $\alpha$ and $\beta$ are zero;

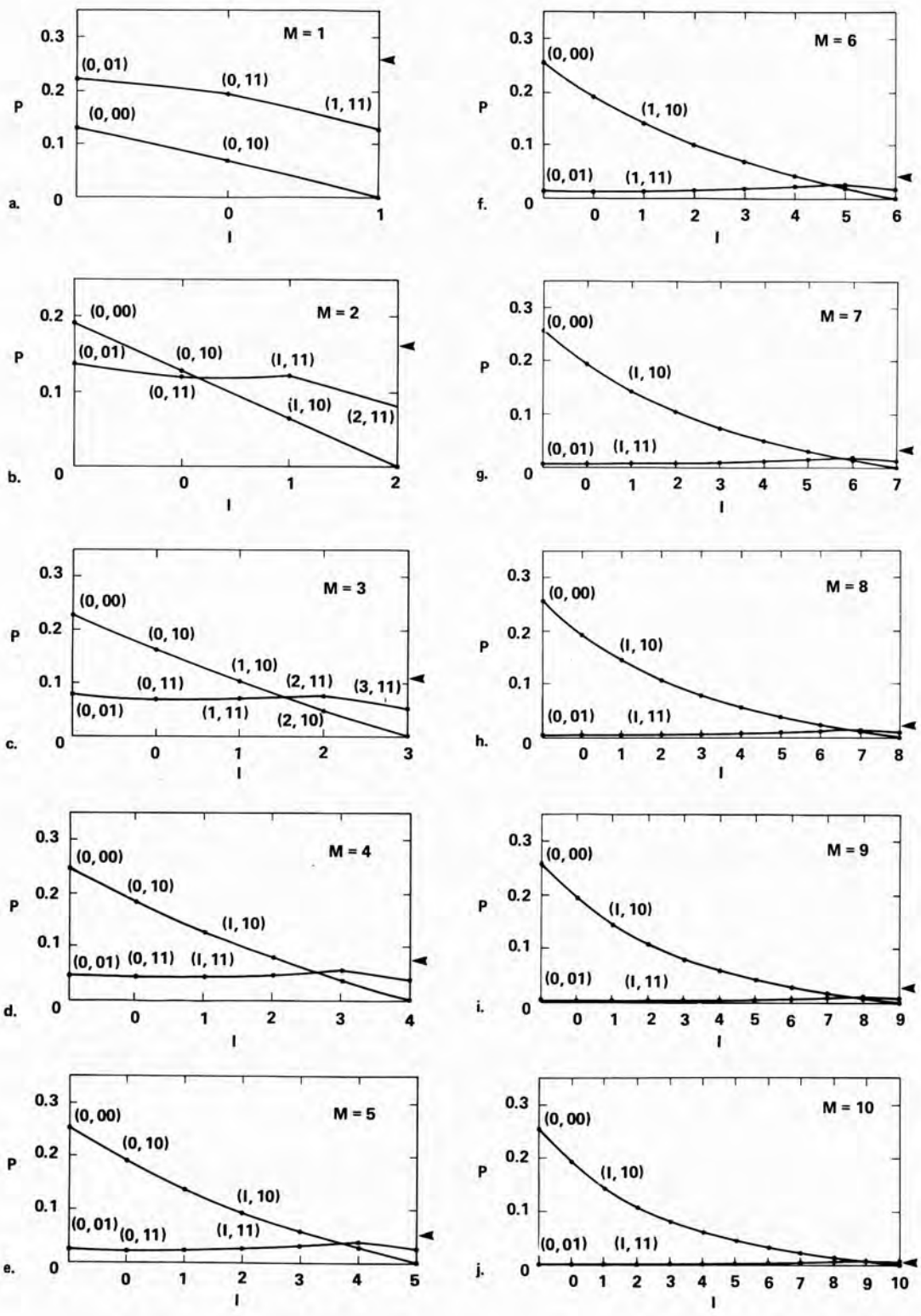



Figure 4.3-1. State Probabilities for $\lambda = 6$, $\mu_1 = 8$, $\mu_2 = 1$, and M = 1 through 10

no probabilistic server invocation is employed). The configuration portrayed is of a system in which $\mu_1 = 8\mu_2$ and $\lambda = \frac{2}{3}(\mu_1 + \mu_2)$.

The ordinate of the graphs of figure 4.3–1 is the probability $p_{q,n_1n_2}$. Each point is labelled with its state designation. On the right ordinate of each graph is a left pointing arrow which indicates the residual amount of probability unaccounted for by the graph, i.e., $\sum_{i=0}^{\infty} p_{M+i,11}$.

Figure 4.3–2 displays $\overline{T}$ as a function of $M + \gamma$, where $\alpha = \beta = \gamma$ (the probabilistic coefficients of $\overline{N}$ as given by eqn 4–3). Note that this is one way of interpretting a real-valued threshold M. Figures 4.3–3 and 4.3–4 enlarge upon selected portions of figure 4.3–2 to show detail more clearly. The singularly striking feature of these figures, most evident in figure 4.3–4, is the cusp–like segments which comprise the figure. This graphically portrays the result of theorem 4.1–1, that probabilistic decision rules are sub–optimal.

In figure 4.3–5, the behavior of the system as a function of $\lambda$ is displayed for fixed (integer) values of the threshold M. This figure, in a graphically complimentary manner to that used before (figure 4.2–4), portrays the mechanism by which M changes as a function of $\lambda$, and highlights the notion that achieving a minimum $\overline{T}$ is accomplished via an envelope of constant–M curves. It is clear from this figure that for low arrival rates, optimal performance is achieved with a large M, while lower values of M improve system performance for higher arrival rates.

Figure 4.3–6 displays the general performance of threshold queueing for a wide variety of server configurations under loading conditions from light to saturation. Superimposed on the diagram of the M* regions are lines of constant $\overline{N}$. We see, for example, from this diagram that we can get nearly the same performance ($\overline{N} \approx .4$) out of two servers of rates 3.5 and .5 as we can from two servers of rates 2.5 each.

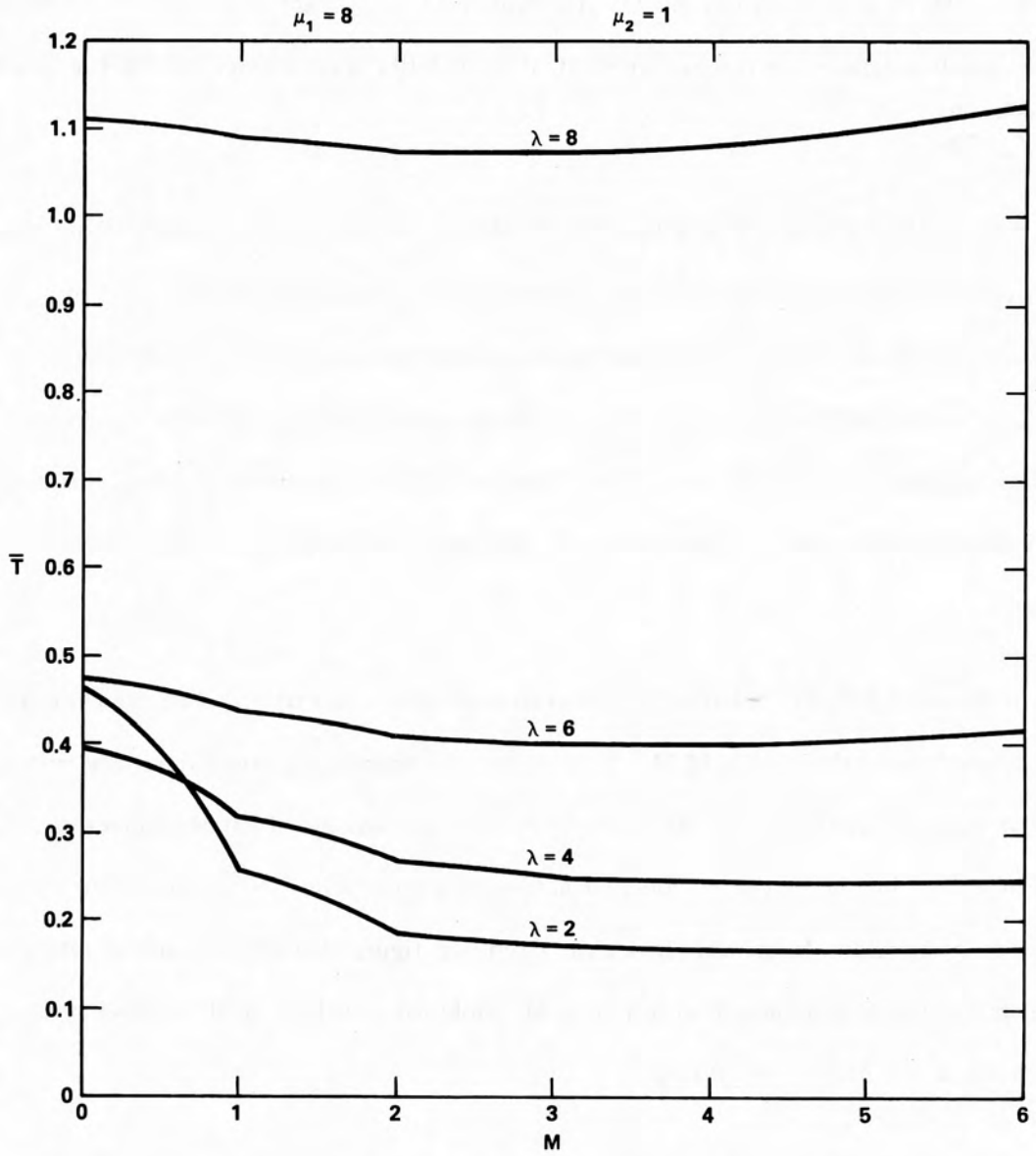



Figure 4.3-2. Mean Time in System as a Function of M

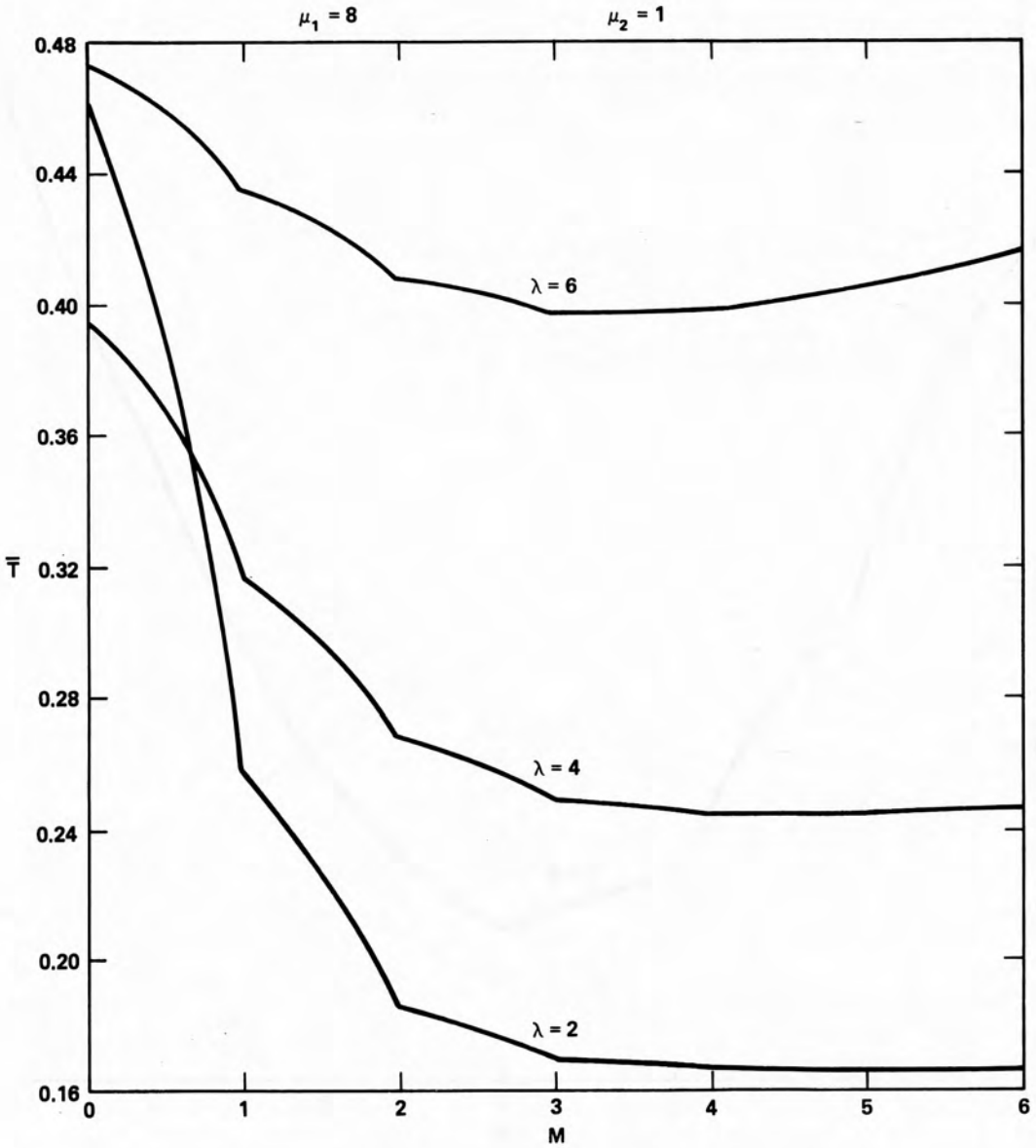



Figure 4.3–3. Mean Time in System as a Function of M (Detail 1)

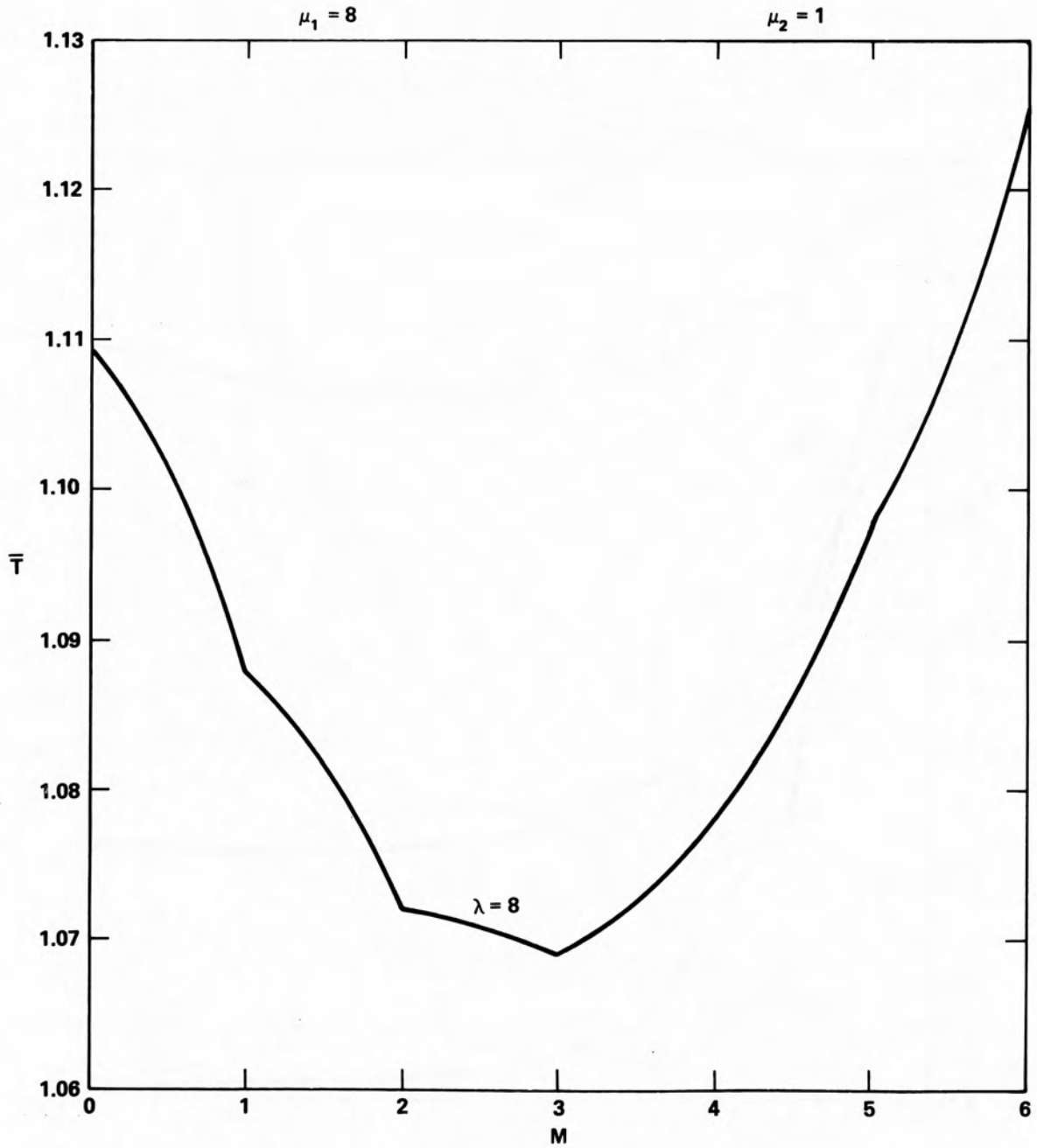



Figure 4.3–4. Mean Time in System as a Function of M (Detail 2)

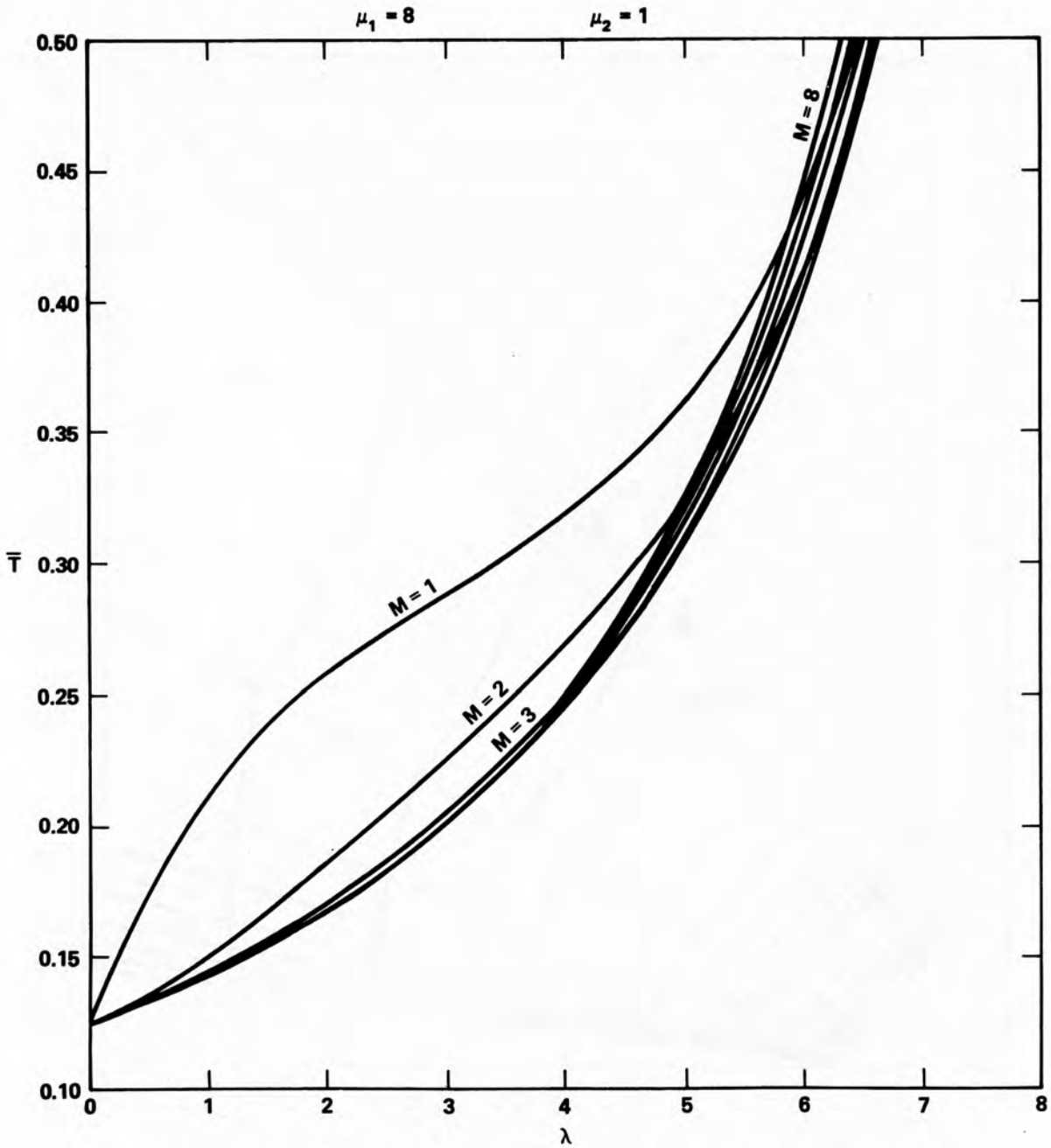



Figure 4.3-5. Mean Time in System as a Function of Arrival Rate for Fixed M

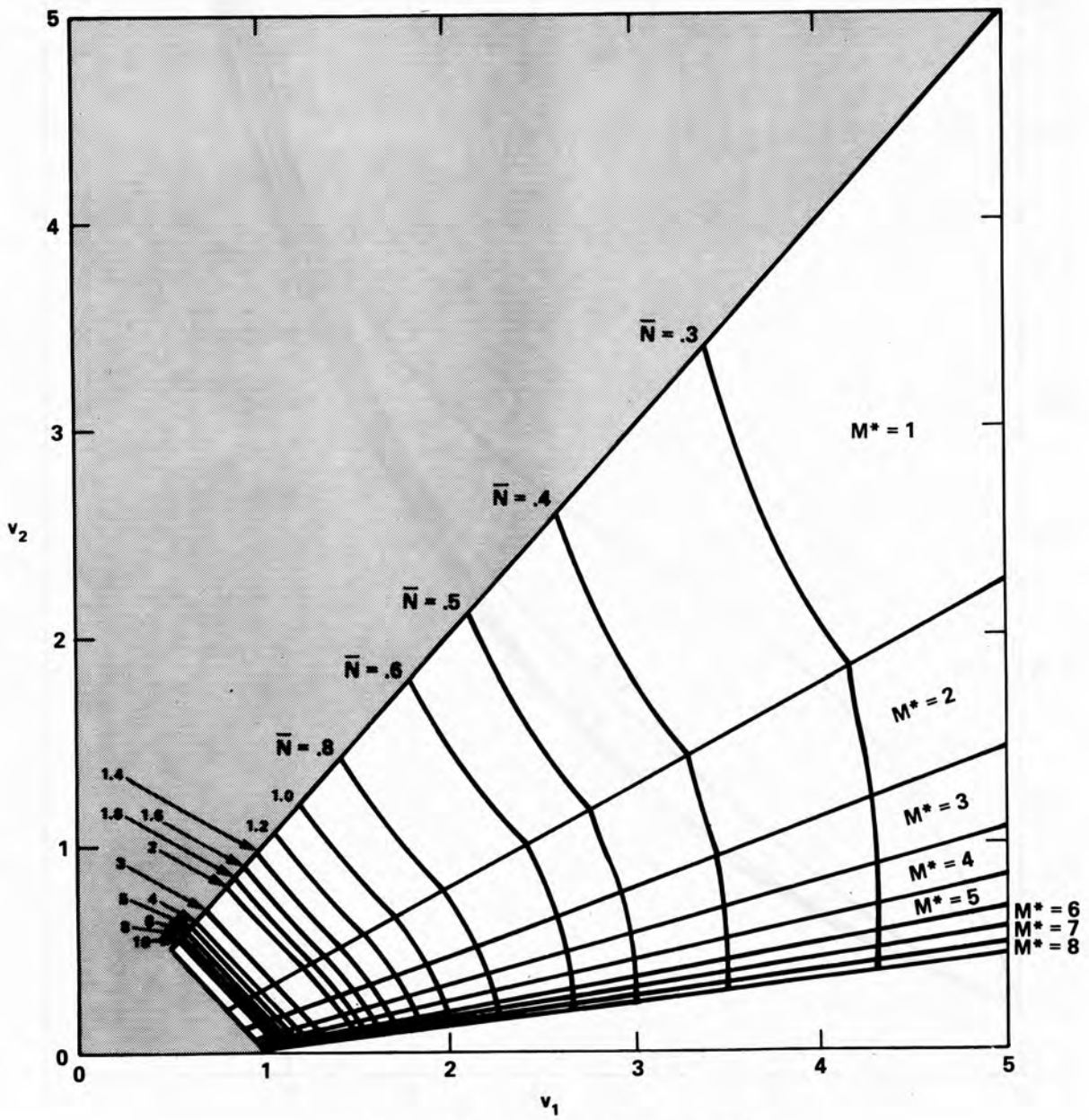



Figure 4.3–6. Lines of Constant $\bar{N}$

A comparison of the threshold queueing optimal performance as compared to two other multi-server queueing disciplines is shown in figure 4.3-7. The graph is a plot of mean time in system as a function of arrival rate for various faster server service rates. The curves labelled "Prob" display the performance of the probabilistic load sharing algorithm derived in [CHENP73]. The curve labelled "Opt-M" displays the performance of the threshold queueing service discipline, and the "Ld-Dep" curves display the performance of the load-dependent queueing discipline. This third discipline is a preemptive discipline, the solution of which is derived in Appendix F. It represents the performance limit of a non-preemptive multiple server system. It is seen from figure 4.3-7 that the threshold queueing discipline performs much like the probabilistic load-sharing discipline under light loading, and approaches the performance of the load-dependent discipline under heavy loading. It is easy to see why this should be the case. The probabilistic load-sharing discipline is adaptive in the sense that the server selection probabilities are a function of the arrival rate, and in this two server case, for a low arrival rate, the selection probability will be 1.0 for the fast server and 0.0 for the slow server, i.e., all customers will be sent to the fast server. In the threshold queueing case, this same effect is created by utilizing a large M*, diverting customers from the slow server.

Under heavier loading, threshold queueing utilizes the slow server more in response to the transients in the system than the probabilistic load sharing discipline, and, hence, outperforms it. To see this more clearly, in the probabilistic case, each arrival is treated independently and sent to a specific server with a fixed probability, regardless of the state of the system. This creates a situation in which an arrival to an empty system may select the slower of the two servers. In the threshold queueing system, this can never happen. A customer is sent to the slower server only after a queue has built up for the fast server.

The performance discussion thus far has focussed on the selection of an optimal M as a function of the service and the arrival rates. From figure 4.3-5, however, it appears

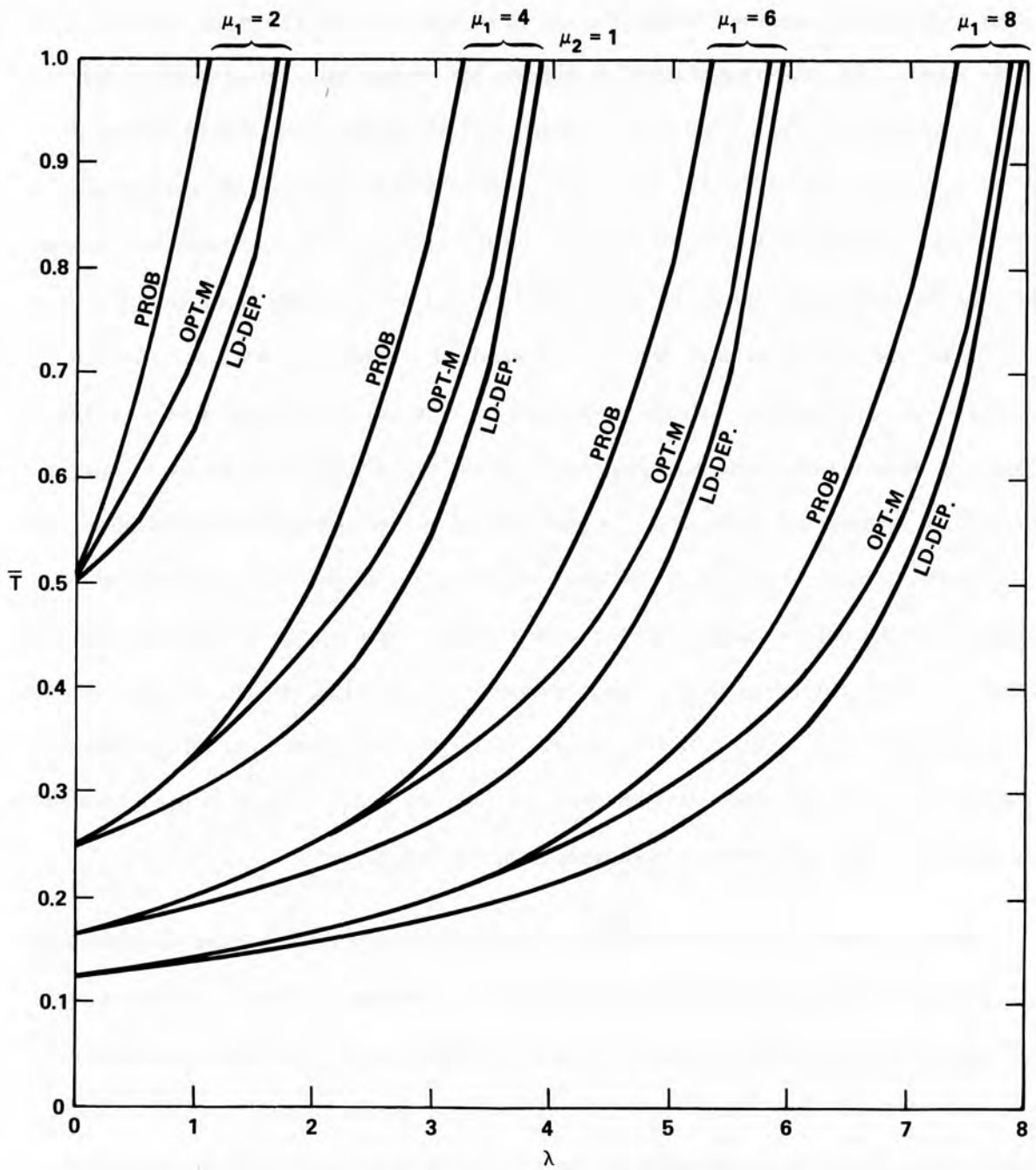



Figure 4.3–7. Performance Comparison

that some of the fixed M curves may come reasonably close to the optimal M curve constructed piece-wise from the fixed M curves. The two most likely candidates for the $\frac{v_1}{v_2} = 8$ case appear to be the M = 3 and the M = 4 curves. Figure 4.3-8 shows the M = 3 curve, which is seen to be not too bad, and, in fact, is the optimal curve for heavy loading. Figure 4.3-9 shows the comparison between the fixed M curve for M = 4 and the optimal M curve, and, on this scale, the difference is barely discernible. In the mid-range of $\lambda$, M = 4 is optimal. For very light loading, the difference between the two curves could hardly be called significant, but the heavy loading range may deserve closer scrutiny since the slope of the performance curve is quite high. Figure 4.3-10 displays the fixed M = 4 curve along with the optimal M (M = 3) curve in the heavy loading range, and the performance difference between the two curves can be seen to be on the order of 1%. The conclusion would seem to be, therefore, that a prudent selection of a fixed M will suffice in all but the most stringent of situations, avoiding the complexity of adjusting the threshold level (M) dynamically as a function of arrival rate.

For those situations requiring dynamic control of the threshold size, eqn 4.2-50 yields a good approximation to M*, and is repeated here:

$$M^* \approx \tilde{M} = \left\lfloor \frac{v_1 - 1 + \sqrt{4v_2 + (v_1 - 1)^2}}{2v_2} \right\rfloor \qquad (2)$$

In figure 4.3-11, the degree to which this approximation to M* affects $\bar{N}$ is displayed. This graph illustrates the worst case effect on $\bar{N}$. Referring to figure 4.2-2, the plotted quantity is $\left[ \frac{\bar{N}(\tilde{M}, v_1, v_2)}{\bar{N}(M^*, v_1, v_2)} - 1 \right] \times 100$ along the $\tilde{M}$ region boundaries, where the maximum displacement from the optimal occurs. It is clear from this figure that the penalty due to this approximation of M* never exceeds 1.5%.

We conclude this section by considering the performance improvement realized by threshold queueing with approximate thresholds $\tilde{M}$ over threshold queueing with fixed M = 1. Fixing M at 1 corresponds to a non-preemptive discipline in which servers are

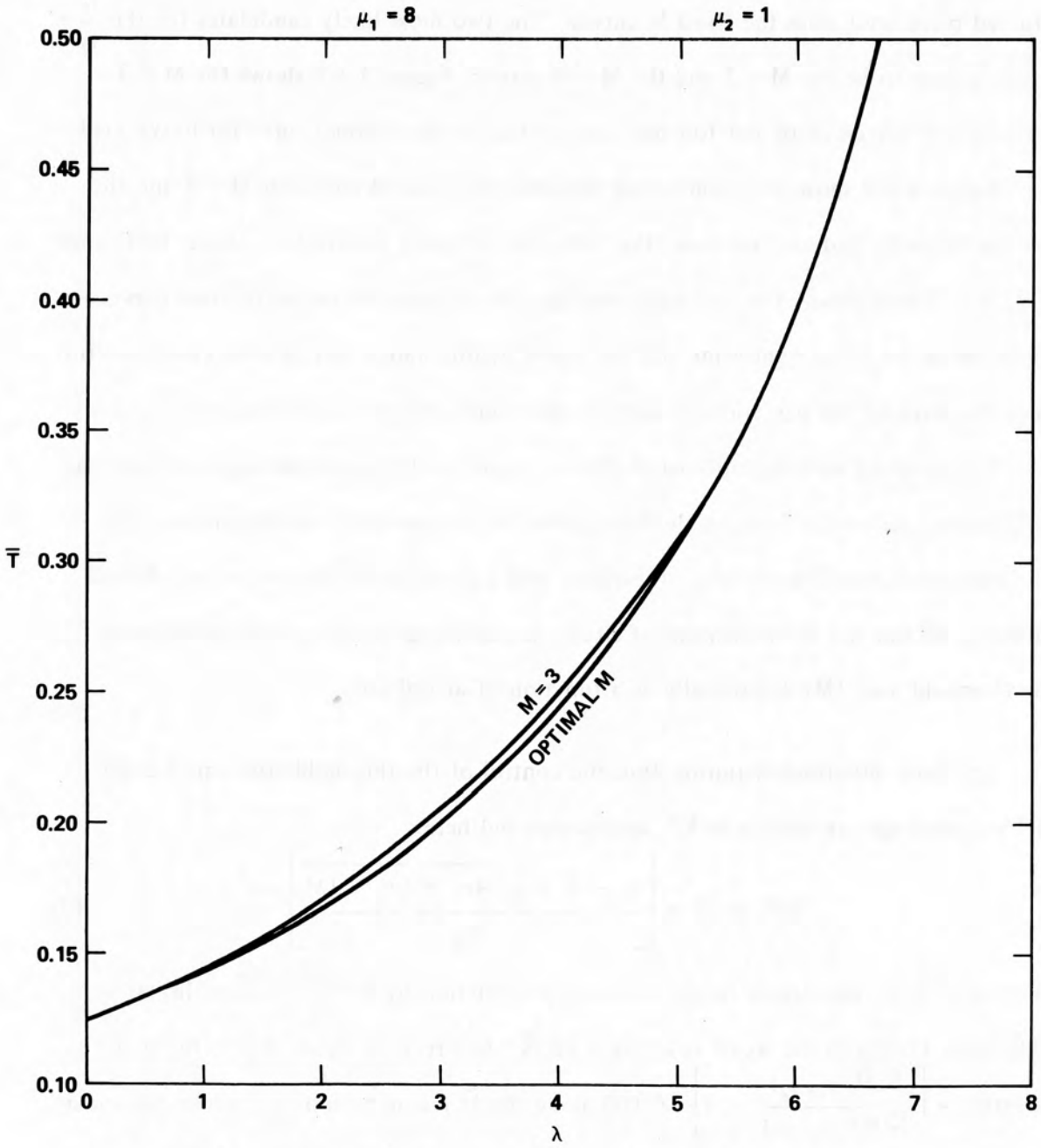



Figure 4.3–8. Using Fixed M = 3 to Approximate Optimal M

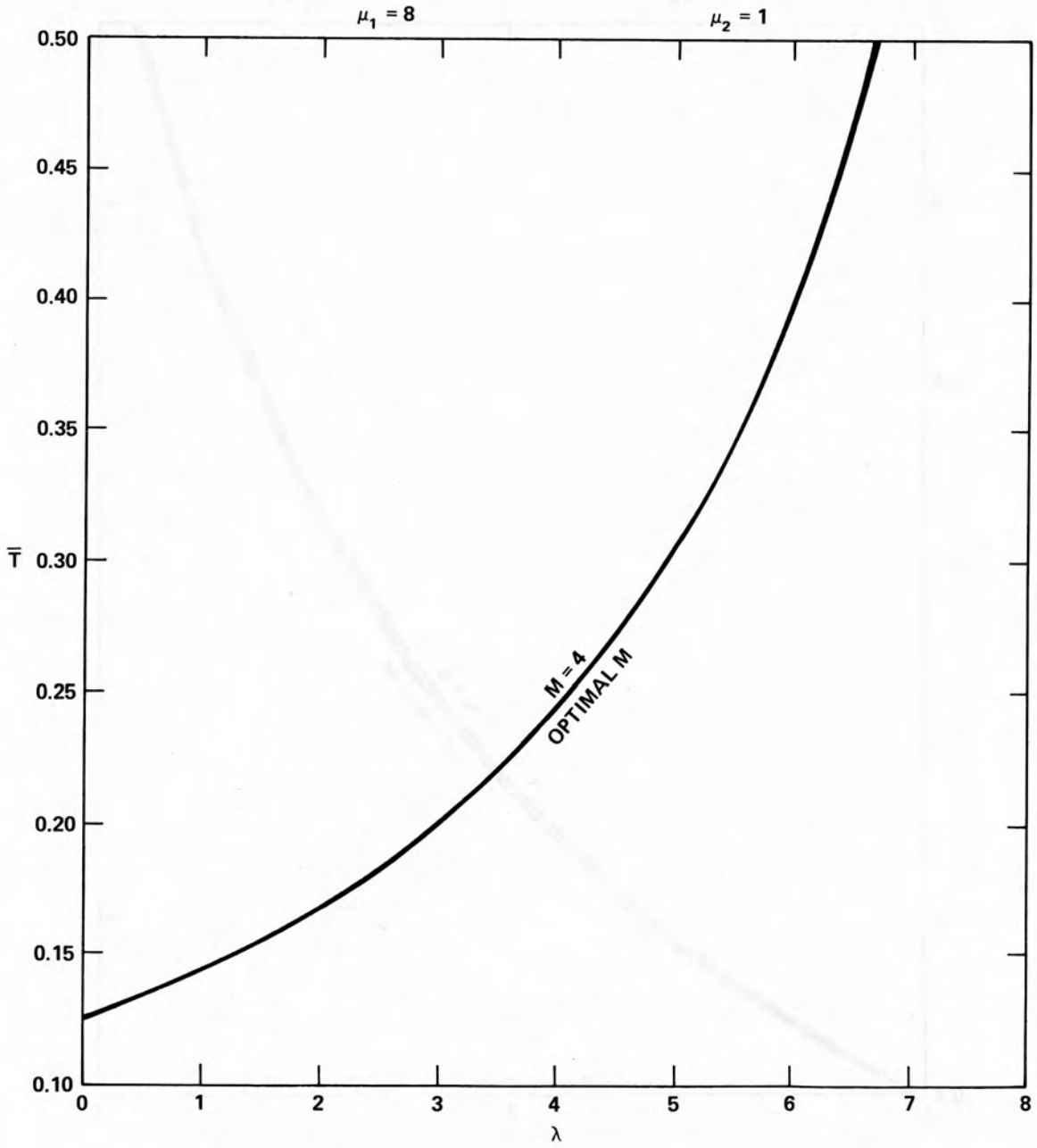



Figure 4.3–9. Using Fixed M = 4 to Approximate Optimal M

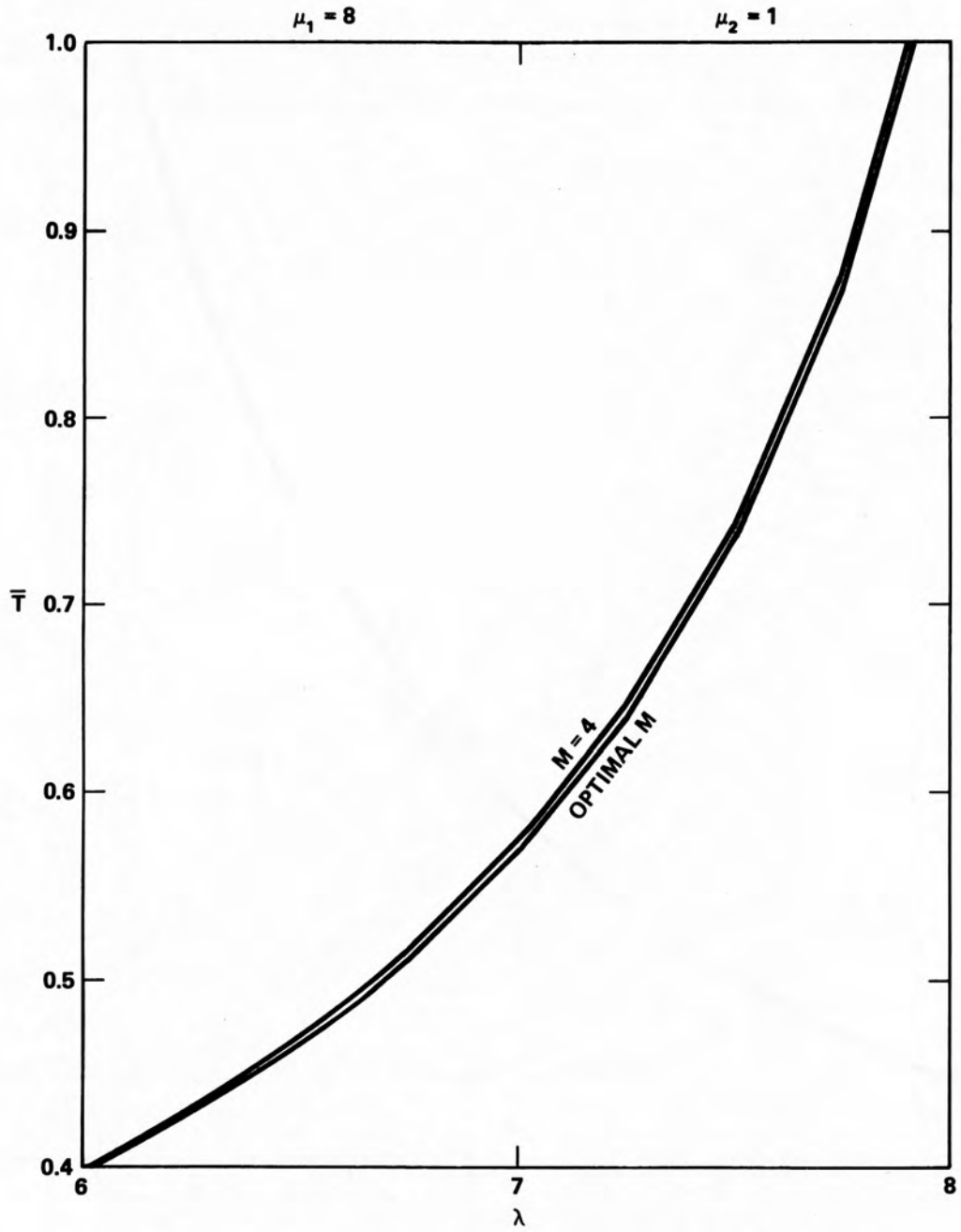



Figure 4.3-10. Using Fixed M = 4 to Approximate Optimal M (Detail)

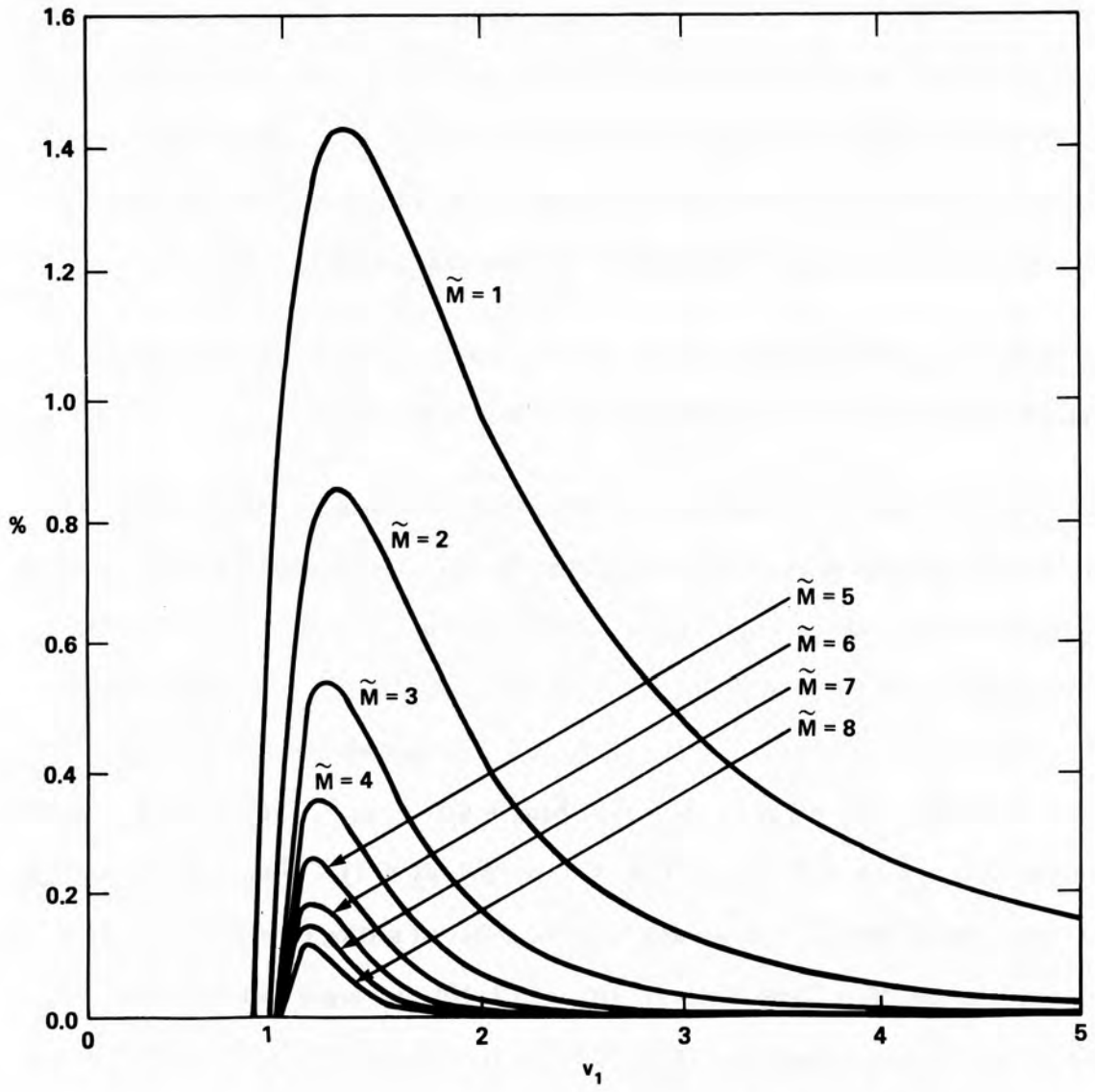



Figure 4.3–11. Maximum $\bar{N}$ Sub–Optimality Due to $\tilde{M}$ Approximation

never left idle if customers are waiting in the queue, with preference given to the faster server. Figure 4.3–12 displays contours of constant performance improvement in the $(v_1, v_2)$ plane.

## 4.4 Conjecture on Probabilistic Server Invocation

It was shown in theorem 4.1–1 that probabilistic thresholds are sub–optimal, and seen in section 4.3 that $\bar{N}$ expressed as a function of the threshold M is concave up with a (quasi) unique optimum threshold. We consider now the more general two server discipline described in 3.2 and illustrated in figure 3.2–5. In this discipline, probabilistic decision rules are allowed for each arrival. We make the following conjecture.

Conjecture 4.4–1 Probabilistic decision rules for server invocation are, in general, sub–optimal in the two server case with $\bar{N}$ as the objective function.

Rationale We begin by considering a modified threshold queueing discipline in which probabilistic decision rules are allowed up through states (M, 10) and (M, 11). This can be realized within the context of figure 3.2–4 by setting $\alpha_i = 0$, $\beta_{i+1} = 0$ for $i > M$. Assume $M > 2$. It has been shown previously that if $\alpha_i = 0$, $\beta_{i+1} = 0$ for $i > 0$, then $\bar{N}(\alpha_0 = \beta_1 = 1) < \bar{N}(\alpha_0 < 1, \beta_1 < 1)$. It has also been shown that if $\alpha_i = \beta_{i+1} = 0$ for $i > 1$, then $\bar{N}(\alpha_0 = \beta_1 = 1, \alpha_1 = \beta_2 = 1) < \bar{N}(\alpha_0 = \beta_1 = 1, \alpha_1 < 1, \beta_2 < 1)$. We question whether it is possible that $\bar{N}(\alpha_0 < 1, \beta_1 < 1, \alpha_1 \leq 1, \beta_2 \leq 1) < \bar{N}(\alpha_0 = \beta_1 = 1, \alpha_1 = \beta_2 = 1)$. Referring to figure 3.2–4, we note that the effect of setting $\alpha_0 = \beta_1 = \alpha_1 = \beta_2 = 1$ is to maximize the "flow" into states (1, 10) and (2, 10), and argue that any attempt at routing system flow toward (1, 11) or (2, 11) at the expense of (1, 10) and (2, 10) will result in decreased system performance (increased $\bar{N}$). This argument can be continued in a step–wise fashion up to the threshold M, leading to the conclusion that if $\alpha_i = \beta_{i+1} = 0$ for $i > M - 1$, then $\alpha_i = \beta_{i+1} = 1$ for $i \leq M - 1$.

The second half of the argument proceeds likewise, but coming down towards M, rather than up to it. It seems reasonable that for operation near system saturation, both

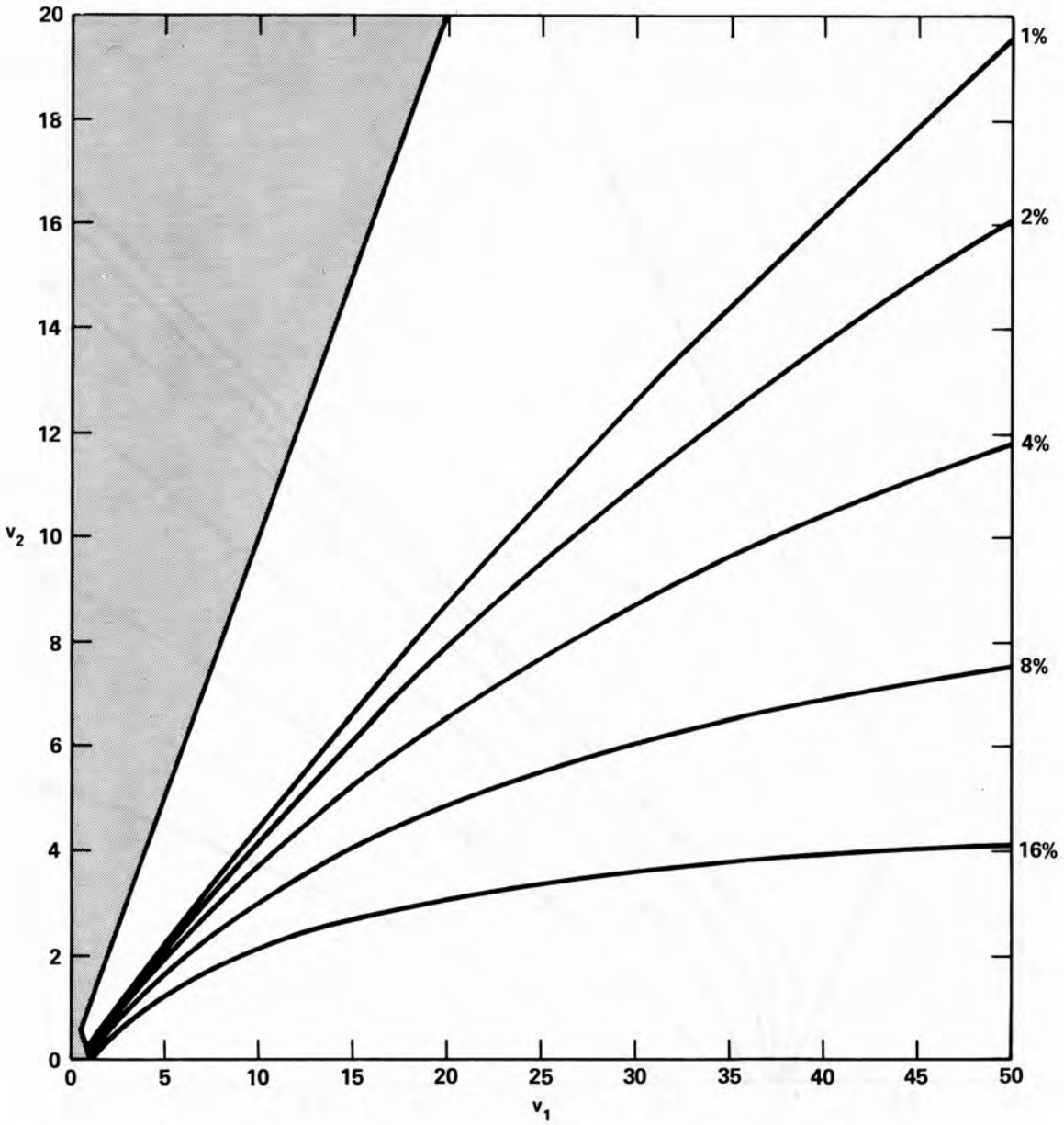



Figure 4.3-12(a). Performance Improvement with Threshold Queueing

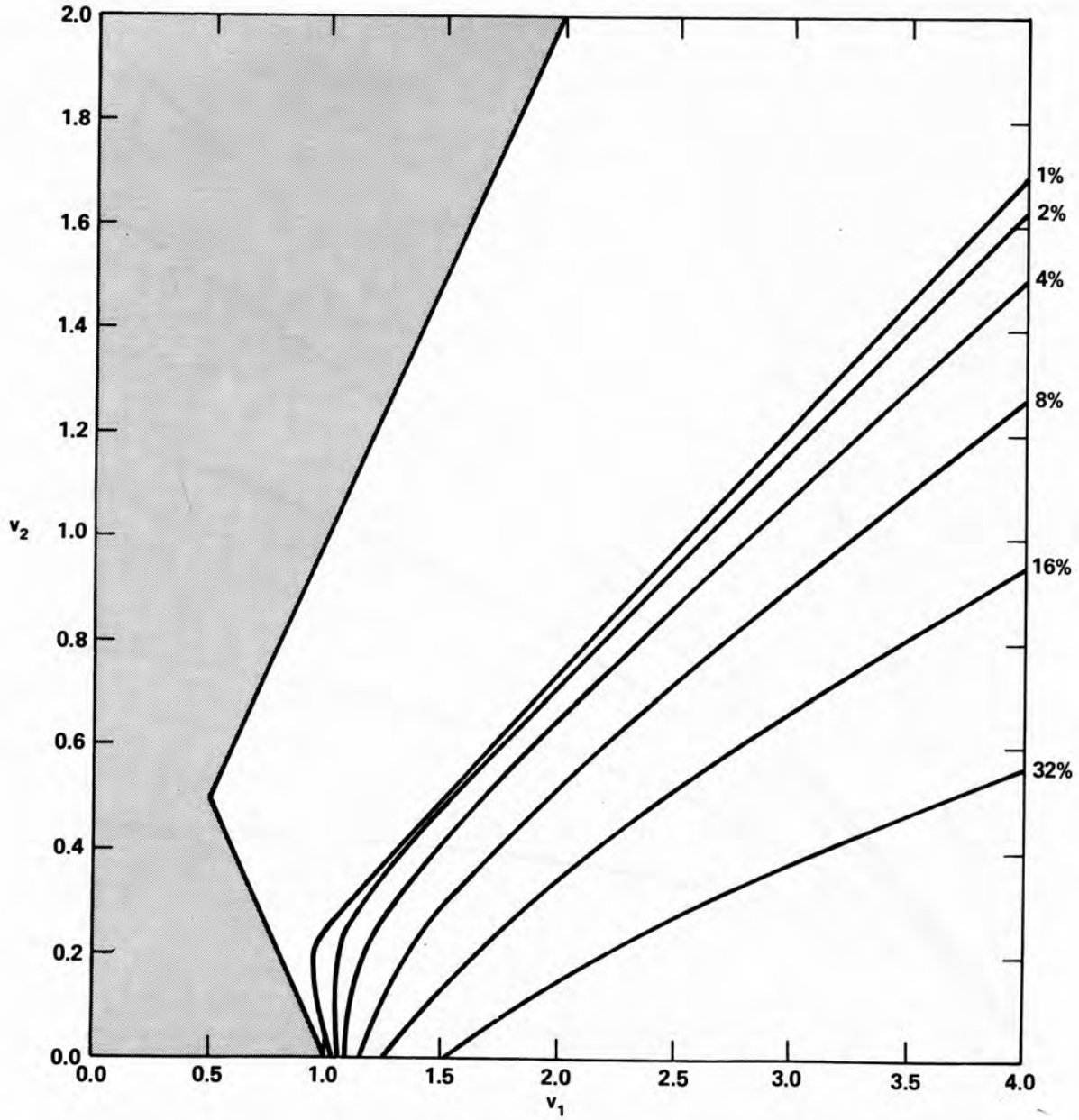



Figure 4.3-12(b). Performance Improvement with Threshold Queueing (Detail)

servers should be in full operation, i.e., that there exists an $\ell$ such that $\alpha_i = \beta_{i+1} = 0$ for $i > \ell$. If we proceed in a stepwise fashion from $\ell$ down towards M, we again encounter a monotonically improving $\overline{N}$, in which system flow is specifically directed away from states (i, 10) for $i > M$.

These arguments coupled with the observation of the distinct cusp-like behavior of $\overline{N}$ for probabilistic thresholds leads to the conjecture.

With strong evidence supporting the viability of this conjecture, the following conjecture is introduced:

Conjecture 4.4-2 Probabilistic decision rules for server invocation are sub-optimal in the n-server case with $\overline{N}$ as the objective function.

Rationale This conjecture follows from conjecture 4.4-1.

Conjecture 4.4-2 leads one to conclude that threshold queueing is to be preferred not only in the two-server case, but also in the n-server case. Whereby for two servers it was sufficient to identify one threshold M, when n servers are considered, n-1 thresholds $M_1, M_2, \ldots, M_{n-1}$ must be determined.

# CHAPTER 5

# THRESHOLD QUEUEING EXTENSIONS AND APPLICATIONS

In the preceding chapters, the focus has been on the analysis and optimization of threshold queueing as a control discipline for two exponential servers of dissimilar speeds which share one queue of infinite capacity. The primary objective has been to minimize the mean number of customers in the system, which resulted in a discipline which preferences the faster of the two servers, often leaving the slow server idle while customers wait. In this chapter, two applications of this discipline are considered. Subsequently, the scope of threshold queueing is extended to consider different objectives and more than two servers. An application is considered in which the slower of the two servers is the preferred server, and consideration is given to utilizing threshold queueing in a system with finite queue capacity to relieve transient overload conditions. The final section of this chapter treats the N-server case.

## 5.1 Applications of Threshold Queueing

Threshold queueing is an appropriate scheduling discipline in any system of multiple, functionally equivalent servers which may operate at dissimilar rates. While potential applications abound, we consider two in this section, one in the control of communications systems, and the other in the scheduling of line printers in a multiprogramming computer system.

Contemporary digital communication networks offer a spectrum of transmission rates, from 110 bits per second (bps) through 1.544 million bps (Mbps) [MARTJ78]. The Bell System's Dataphone Digital System (DDS), for example, offers rates of 2.4, 4.8, 9.6, and 56 thousand bps (Kbps). The performance of communication networks can be analyzed using queueing models in which messages correspond to customers and communication links to servers. Service rates ($\mu$) are measured in units of messages per second, and are a function of message length ($\ell$) and transmission rate ($r$) according to the formula $\mu = r/\ell$.

We will consider using threshold queueing to manage the one-way flow of messages between two cities over two communication links of potentially differing bandwidth (service rate). Figure 5.1-1 displays the ratio of service rates for various combinations of commercially-available communication link bandwidths. We consider a system configuration in which a primary link of 19.2 Kbps is used in conjunction with a secondary link of 1.2, 2.4, or 4.8 Kbps. Figure 5.1-2 displays in tabular form the threshold values over the entire operational range for each of these combinations, where message lengths are assumed to be exponentially distributed with a mean of 1000 bits. These threshold values are computed using the asymptotic approximation given by eqn 4.2-50. Figure 5.1-3 displays the performance as measured by $\bar{N}$ for each of these combinations under the threshold queueing discipline. For comparison, the performance of the primary link operating by itself is shown as predicted by the single server (M/M/1) solution. In figure 5.1-4, the performance improvement realizable by threshold queueing over the performance of a primary link working in conjunction with a secondary link under a threshold discipline with M = 1 is shown. This second discipline corresponds to the intuitive discipline of keeping all links as busy as possible, and invoking them in the inverse order of service rate. The quantity plotted in this figure is $\left[\frac{\bar{N}\left(1, \frac{19.2}{\lambda}, \frac{r}{\lambda}\right)}{\bar{N}\left(\tilde{M}, \frac{19.2}{\lambda}, \frac{r}{\lambda}\right)} - 1\right] \times 100$ where $\bar{N} = \bar{N}(M, v_1, v_2)$, as before, and r = 1.2, 2.4, or 4.8. The discontinuities in the curves occur at points at which $\tilde{M}$ changes. Recalling figure 4.3-12, which is repeated here as figure 5.1-5 with lines of constant $\frac{v_1}{v_2}$ superimposed, corresponding to the three cases just considered, we can readily see in a more general context the performance improvement realizable through threshold queueing. Recognizing that the maximum theoretical performance realizable from a multiple server system is given by the load-dependent solution (Appendix E) and results from a preemptive discipline favoring the faster server, we show in figure 5.1-6 the improvement achievable through such a discipline over threshold queueing. While the preemptive nature of this discipline generally makes it impractical in real systems, it provides a useful

| $r_2$ Kbps \ $r_1$ Kbps | .1 | .3 | 1.2 | 2.4 | 4.8 | 9.6 | 19.2 | 56 | 224 |
|---|---|---|---|---|---|---|---|---|---|
| .1 | 1 | 3 | 12 | 24 | 48 | 96 | 192 | 560 | 2240 |
| .3 | | 1 | 4 | 8 | 16 | 32 | 64 | 187 | 747 |
| 1.2 | | | 1 | 2 | 4 | 8 | 16 | 47 | 187 |
| 2.4 | | | | 1 | 2 | 4 | 8 | 23 | 93 |
| 4.8 | | | | | 1 | 2 | 4 | 11.7 | 46.7 |
| 9.6 | | | | | | 1 | 2 | 5.8 | 23.3 |
| 19.2 | | | | | | | 1 | 2.9 | 11.7 |
| 56 | | | | | | | | 1 | 4 |
| 224 | | | | | | | | | 1 |

Figure 5.1-1. $r_1/r_2 = v_1/v_2$ for standard digital communication

comparison. Again, the discontinuities correspond to jumps in $\widetilde{M}$. It is clear from this figure that the performance of threshold queueing in this application compares quite favorably with the theoretically optimum, with less than 15% further improvement achievable.

For the second example, a reasonably large-scale multiprogramming system is considered. More specifically, we hypothesize the existence of two spooling printers of differing rates. The output from jobs processed by the system is spooled to secondary storage (queued), and upon completion of the job and availability of a printer, the output gets printed. As in the communications example, printer speeds to be considered are representative of those available commercially. Figure 5.1-7 displays ratios of printer speeds (which equal the ratio of service rates) for typical printers currently available. The printed output of jobs corresponds to customers in this system, and the printers to servers. The queue length is the number of jobs whose output awaits printing. Figure 5.1-8 displays threshold size (the queue length at which output is routed to the slower printer)

| $r_1$ (Kbps) | $r_1/r_2$ | Threshold Value (M) | Range of $v_1$ | Message Arrival Rate (1000 bit msgs/sec.) |
|---|---|---|---|---|
| 19.2 | 4 | 1 | .8 - 1 | 19 - 24 |
| | | 2 | 1 - 2.67 | 7 - 19 |
| | | 3 | 2.67 - ∞ | 0 - 7 |
| | 8 | 2 | .89 - 1.07 | 18 - 22 |
| | | 3 | 1.07 - 1.50 | 13 - 18 |
| | | 4 | 1.50 - 2.13 | 9 - 13 |
| | | 5 | 2.13 - 3.33 | 6 - 9 |
| | | 6 | 3.33 - 6.86 | 3 - 6 |
| | | 7 | 6.86 - ∞ | 0 - 3 |
| | 16 | 3 | .94 - 1.00 | 19 - 20 |
| | | 4 | 1.00 - 1.16 | 17 - 19 |
| | | 5 | 1.16 - 1.33 | 14 - 17 |
| | | 6 | 1.33 - 1.52 | 13 - 14 |
| | | 7 | 1.52 - 1.75 | 11 - 13 |
| | | 8 | 1.75 - 2.03 | 9 - 11 |
| | | 9 | 2.03 - 2.40 | 8 - 9 |
| | | 10 | 2.40 - 2.91 | 7 - 8 |
| | | 11 | 2.91 - 3.67 | 5 - 7 |
| | | 12 | 3.67 - 4.92 | 4 - 5 |
| | | 13 | 4.92 - 7.43 | 3 - 4 |
| | | 14 | 7.43 - 14.93 | 1 - 3 |
| | | 15 | 14.93 - ∞ | 0 - 1 |

Figure 5.1-2. Threshold values for communications example

as a function of $v_1$. The corresponding mean arrival rate of output to be printed is also shown, where it is assumed that a typical output listing consists of 1000 lines of printer output. In this example we consider a secondary 300 line per minute printer which is used to augment a faster primary printer. Primary printer rates of 900, 1200, 1600, 2000, and 2400 lines per minute are considered.

Figure 5.1-9 illustrates the performance of these various printer combinations under threshold queueing as a function of the arrival rate $\lambda$ (jobs/sec.) of jobs to be printed. In figure 5.1-10 the performance curves for the M = 1 discipline are superimposed on the threshold queueing curves for comparison. The M = 1 discipline in this case corresponds to keeping both printers busy if possible, and starting the faster one first following idle periods. The performance improvement realized by threshold queueing over the

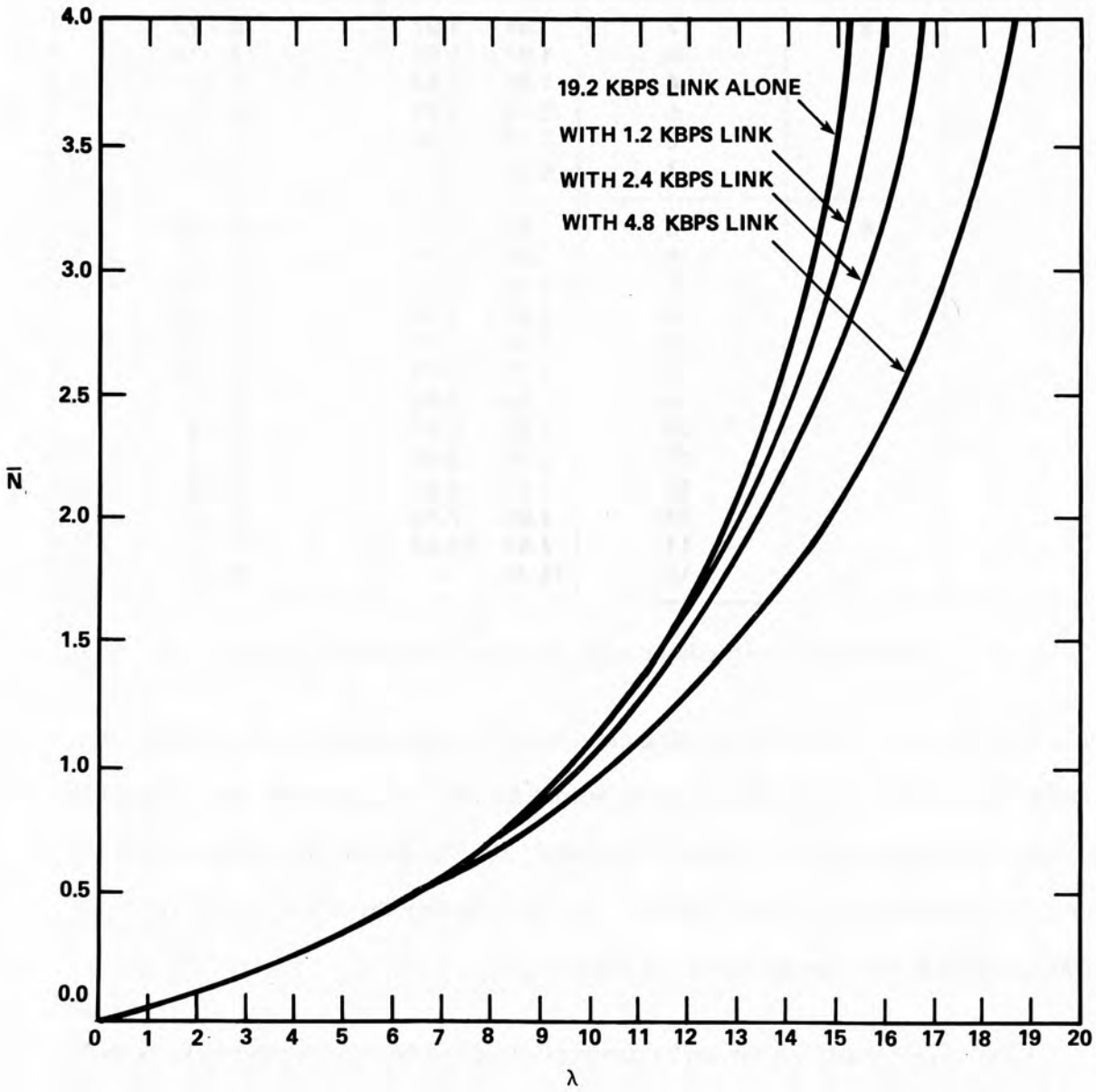



Figure 5.1-3. Performance of a 19.2 Kbps Communications System with a Secondary Link under Threshold Queueing

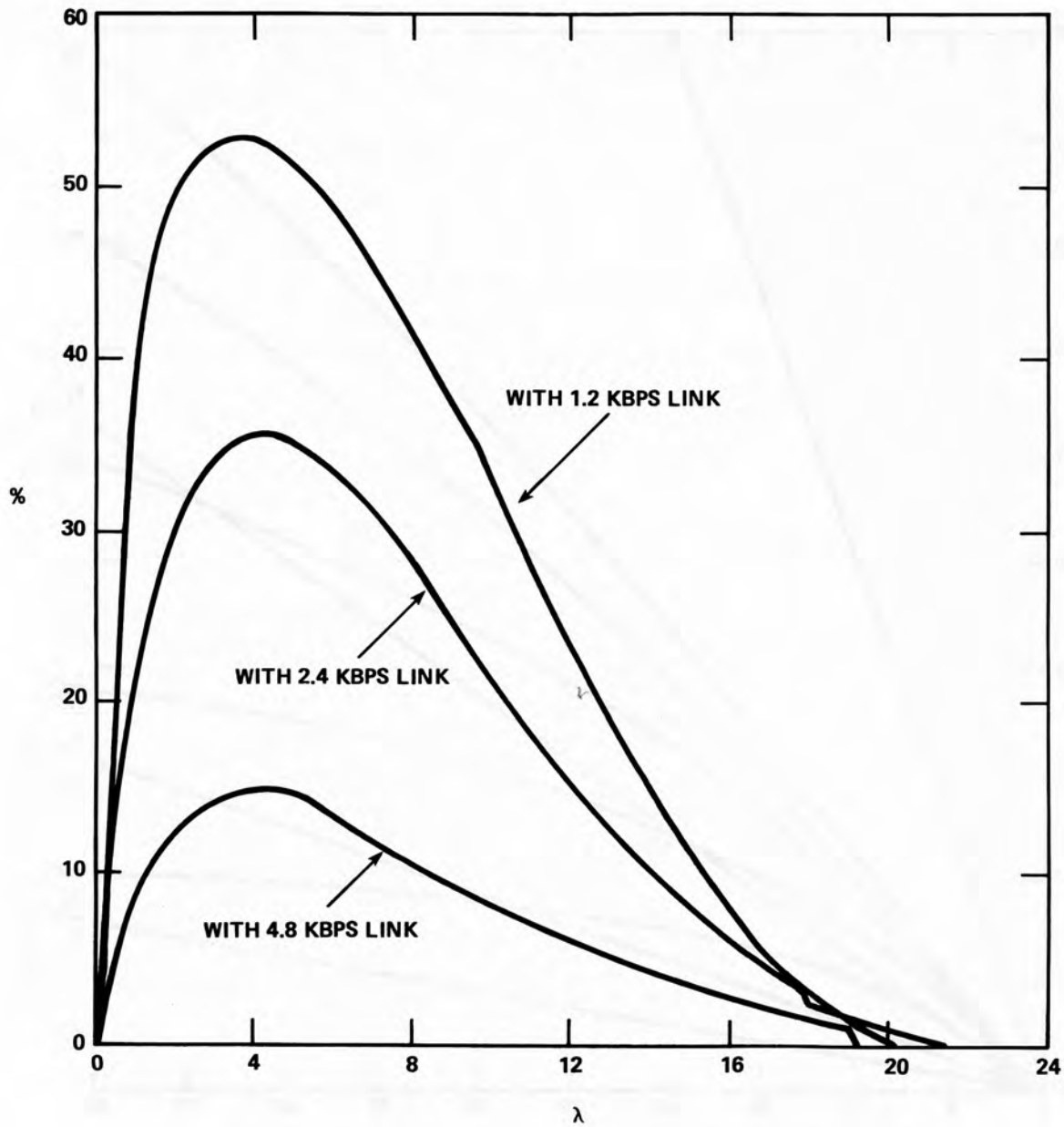



Figure 5.1–4. Performance Improvement of Threshold Queueing over M = 1 Discipline for a 19.2 Kbps Communications System

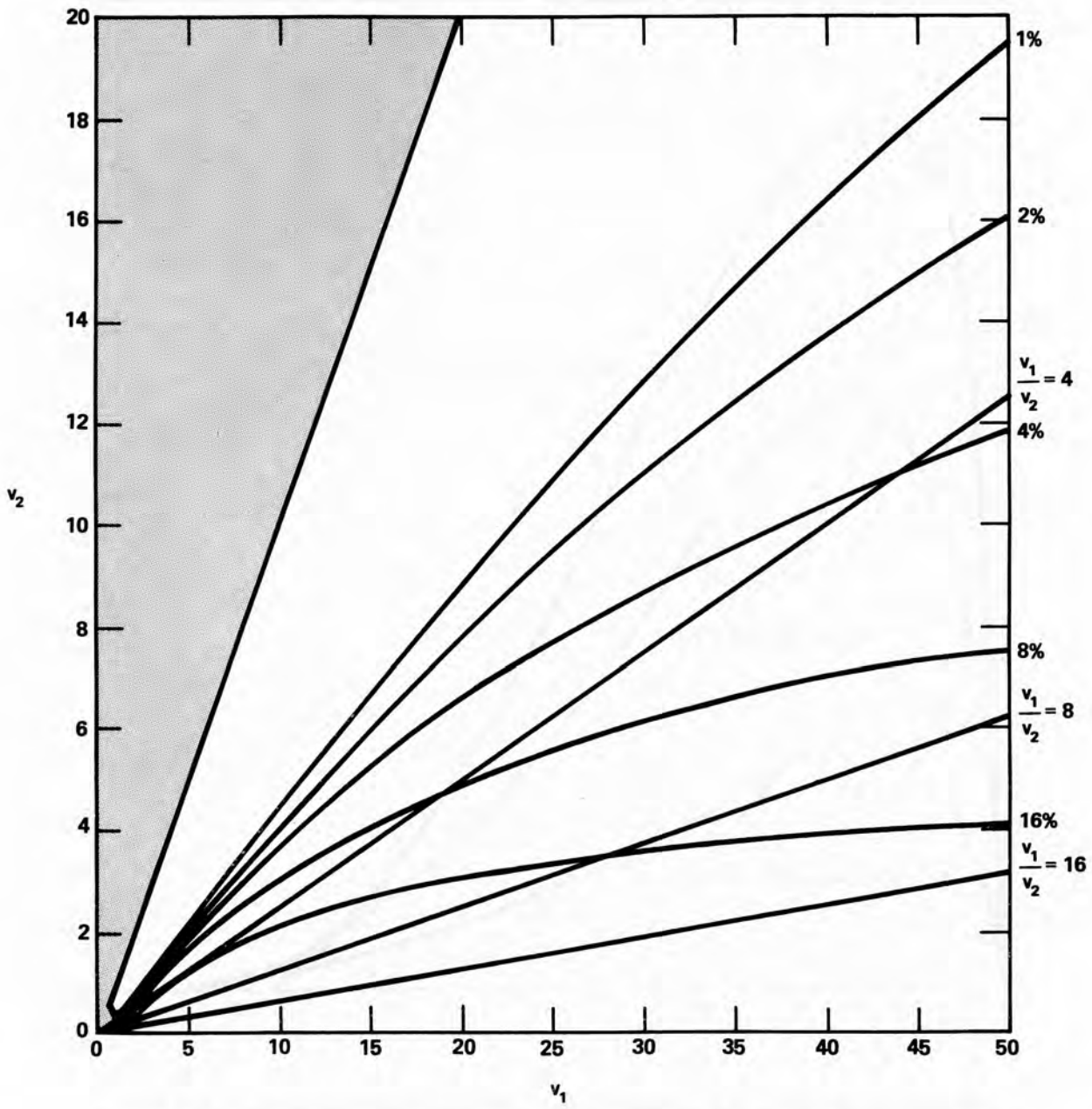



Figure 5.1–5(a). Performance Improvement for Different Ratios of Transmission Rates

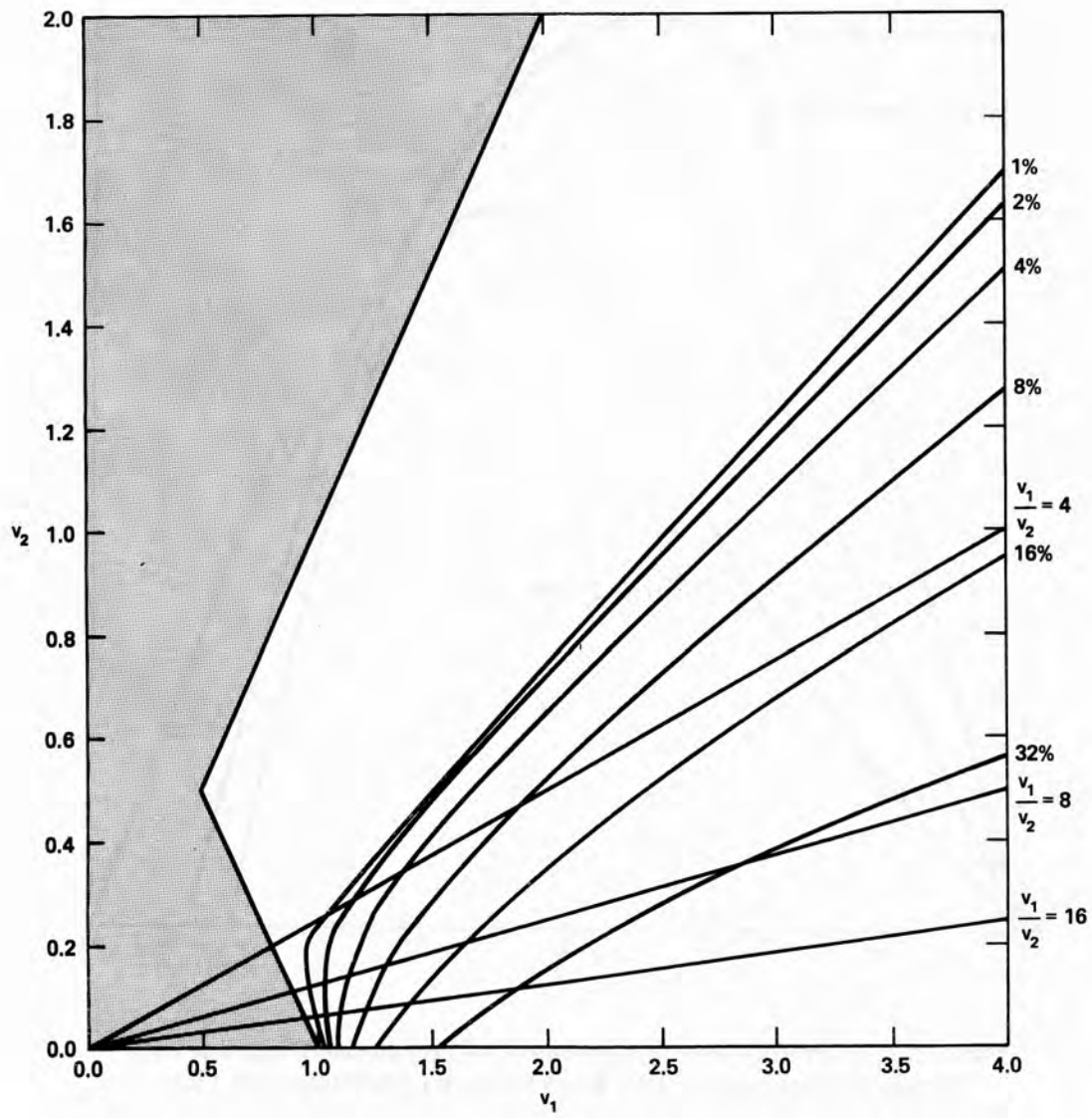



Figure 5.1-5(b). Performance Improvement for Different Ratios of Transmission Rates (Detail)

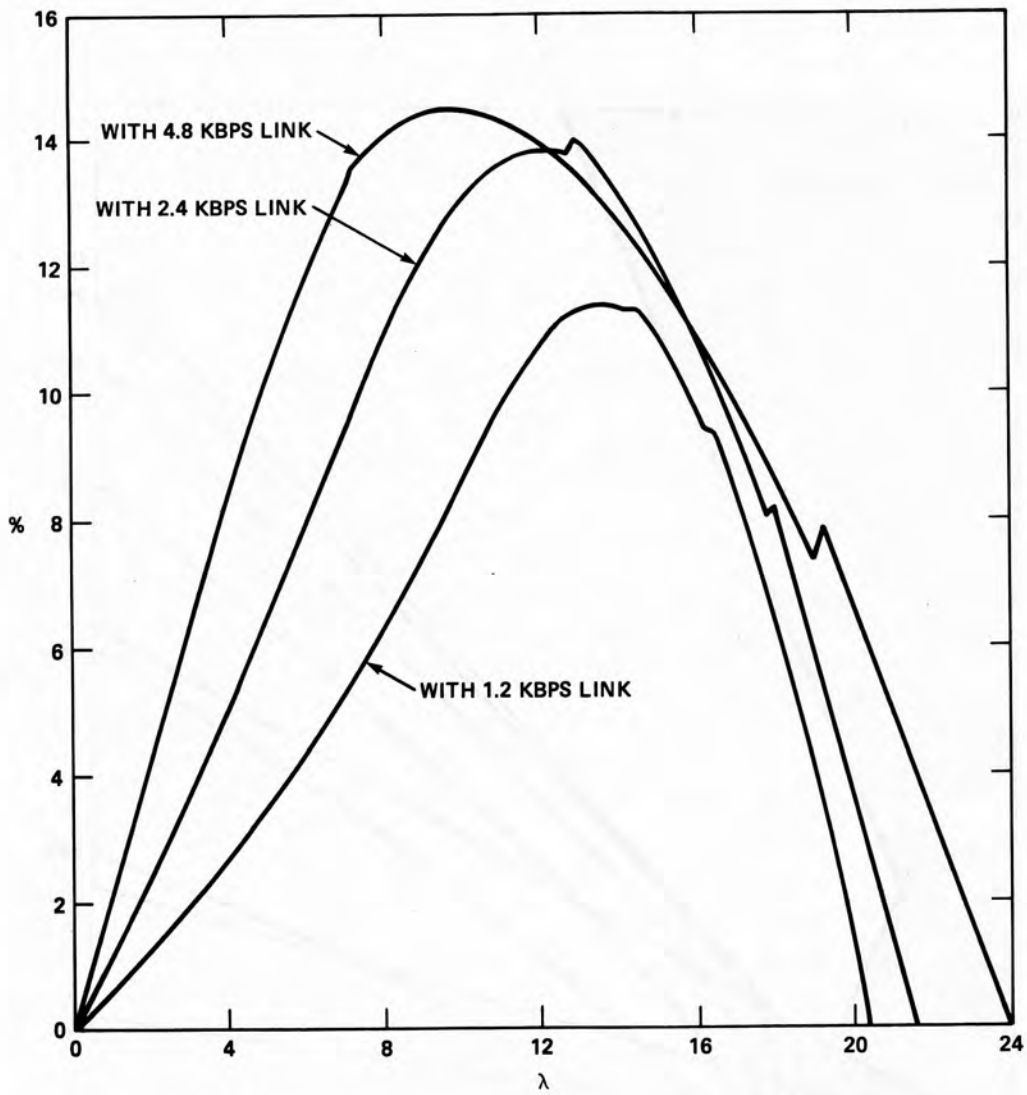



Figure 5.1-6. Performance Improvement of Preemptive Discipline over Threshold Queueing for 19.2 Kbps Primary Communication Link

| $r_2$ lps \ $r_1$ lps | 150 | 300 | 600 | 900 | 1200 | 1600 | 2000 | 2400 |
|---|---|---|---|---|---|---|---|---|
| 150 | 1 | 2 | 4 | 6 | 8 | 10.7 | 13.3 | 16 |
| 300 | | 1 | 2 | 3 | 4 | 5.3 | 6.7 | 8 |
| 600 | | | 1 | 1.5 | 2 | 2.7 | 3.3 | 4 |
| 900 | | | | 1 | 1.3 | 1.8 | 2.2 | 2.7 |
| 1200 | | | | | 1 | 1.3 | 1.7 | 2 |
| 1600 | | | | | | 1 | 1.3 | 1.5 |
| 2000 | | | | | | | 1 | 1.2 |
| 2400 | | | | | | | | 1 |

Figure 5.1–7. $r_1/r_2 = v_1/v_2$ for line printers

M = 1 discipline is displayed in figure 5.1–11, and the performance improvement which would be realized by a preemptive load–dependent discipline over threshold queueing is shown in figure 5.1–12. As was the case with the communications example, depending on the ratio of the two service rates, threshold queueing can improve the overall performance markedly, while further improvement toward the limiting preemptive load–dependent case is bounded at about 15%.

## 5.2 Slow Server Preference

There exist situations in which, either due to constraints or by design, the slower of two servers is the preferred server. As an example, we consider an organization which owns a relatively small (and slow) computer system, but which may buy time on a larger, faster computer system as necessary to provide an acceptable level of service to its users. Rather than deal directly with cost optimization, we consider as an objective maintaining a consistent level of performance to the user community in spite of potentially changing loading conditions. Because of the inversion in server preference, the optimization

| $r_2$ lps | $r_1/r_2$ | Threshold Value (M) | Range of $v_1$ | Job Arrival Rate (1000 line jobs/sec.) |
|---|---|---|---|---|
| 300 | 3 | 1 | .75 - 1.50 | .6 - 1.2 |
| | | 2 | 1.50 - ∞ | 0 - .6 |
| | 4 | 1 | .80 - 1.00 | 1.2 - 1.5 |
| | | 2 | 1.00 - 2.67 | .5 - 1.2 |
| | | 3 | 2.67 - ∞ | 0 - .5 |
| | 5.3 | 2 | .84 - 1.54 | 1 - 1.9 |
| | | 3 | 1.54 - 3.06 | .5 - 1 |
| | | 4 | 3.06 - 14.13 | .1 - .5 |
| | | 5 | 14.13 - ∞ | 0 - .1 |
| | 6.7 | 2 | .87 - 1.21 | 1.7 - 2.3 |
| | | 3 | 1.21 - 1.86 | 1 - 1.7 |
| | | 4 | 1.86 - 3.15 | .6 - 1 |
| | | 5 | 3.15 - 7.98 | .3 - .6 |
| | | 6 | 7.98 - ∞ | 0 - .3 |
| | 8 | 2 | .89 - 1.07 | 2.2 - 2.7 |
| | | 3 | 1.07 - 1.50 | 1.6 - 2.2 |
| | | 4 | 1.50 - 2.13 | 1.1 - 1.6 |
| | | 5 | 2.13 - 3.33 | .7 - 1.1 |
| | | 6 | 3.33 - 6.86 | .4 - .7 |
| | | 7 | 6.86 - ∞ | 0 - .4 |

Figure 5.1-8. Threshold values for line printer example

procedure of section 4.2 does not apply. The objective of slow server preference is not to minimize the mean number of customers in the system, but instead to relieve a transient system overload condition. The typical situation is one in which a slow server attempts to process the incoming stream of customers, but switches in a faster server when the queue builds up to a given threshold to prevent inordinate customer wait times.

While optimization in the sense treated in chapter 4 is inapplicable, the basic threshold queueing formulation of chapter 3 remains valid. The region of the $(v_1, v_2)$ plane which is used for slow server preference is shown in figure 5.2-1. It is that region where $v_2 > v_1$ and $v_2 + v_1 > 1$. Threshold queueing can be used to maintain a degree of control over $\overline{N}$ (and, hence, $\overline{T}$) by selectively invoking the faster server as a function of system loading. This is illustrated in figure 5.2-2 over the performance range of the

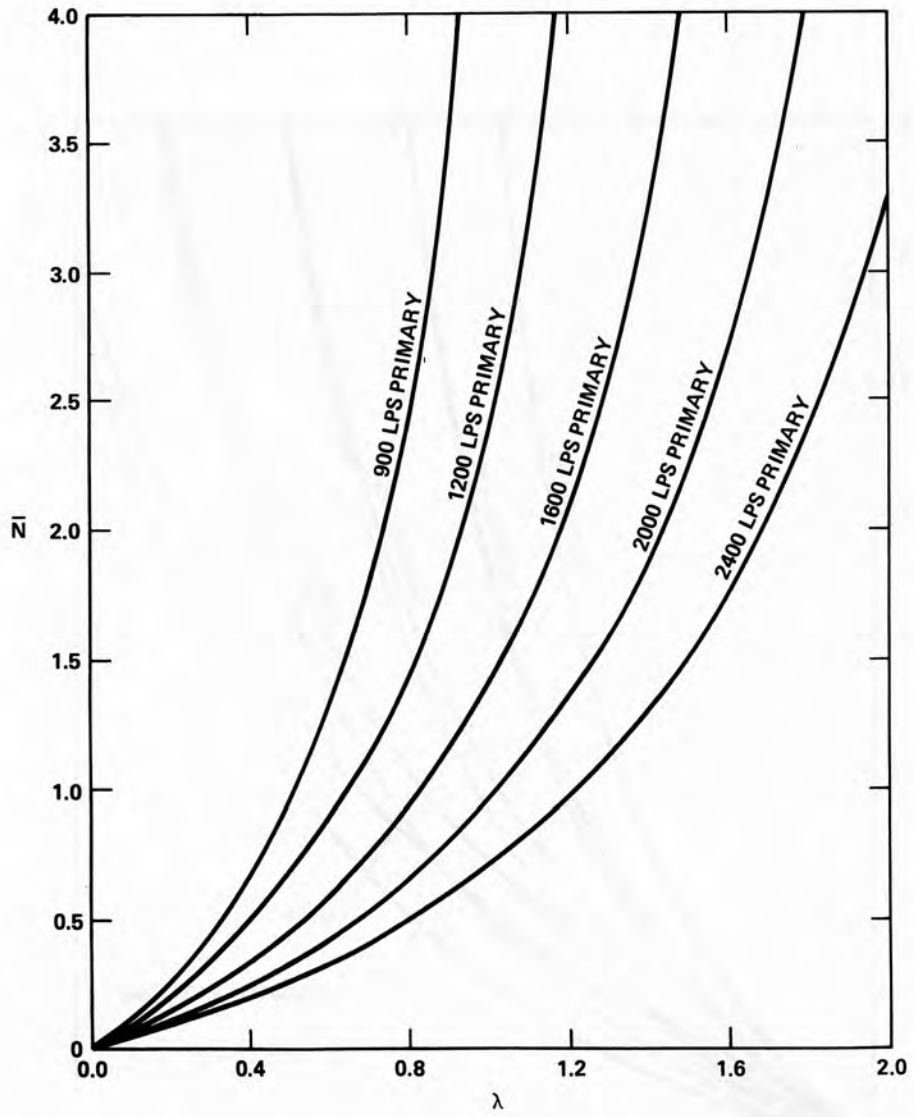



Figure 5.1-9. Performance of Dual Printer Configuration with 300 lps Secondary Printer under Threshold Queueing

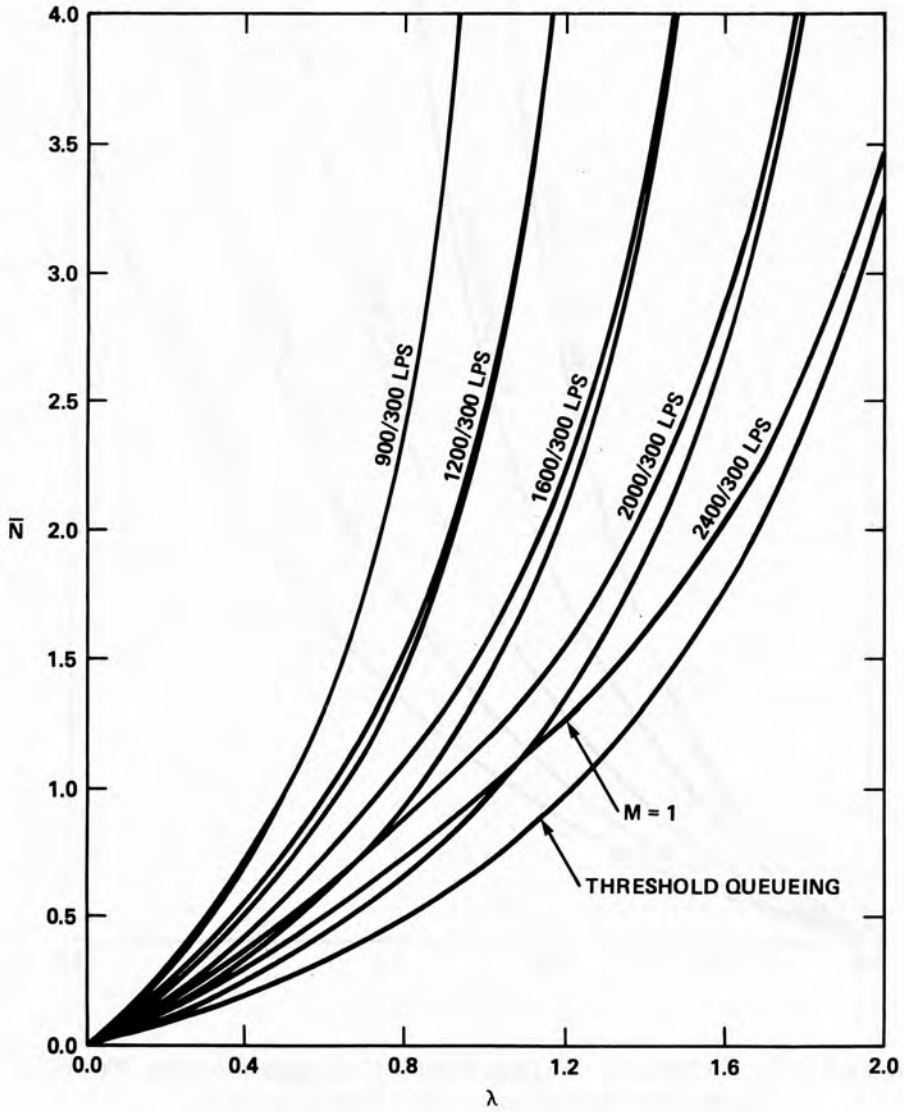



Figure 5.1-10. Comparison of M = 1 Discipline to Threshold Queueing for Line Printers

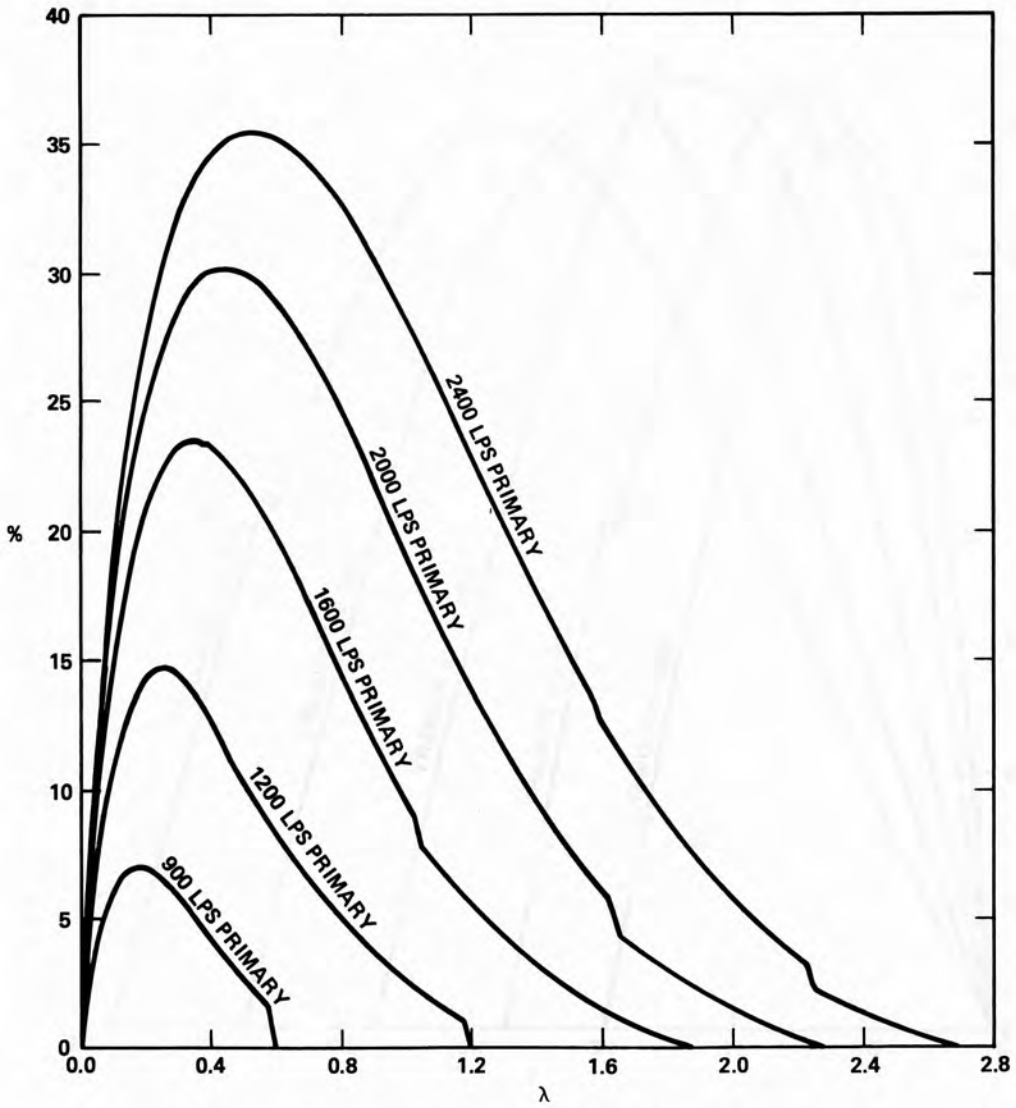



Figure 5.1-11. Performance Improvement of Threshold Queueing over M = 1 for Dual Printers with 300 lps Secondary Printer

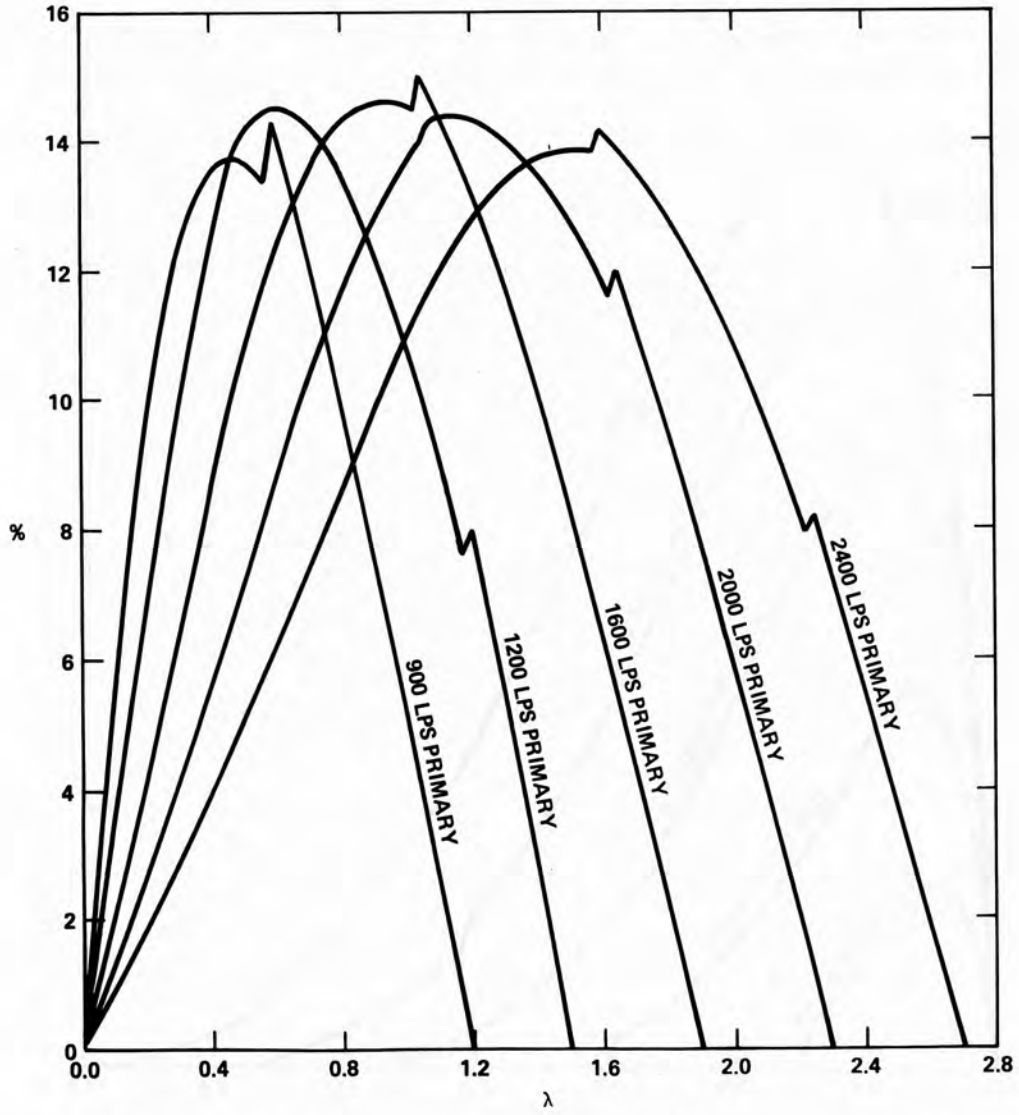



Figure 5.1-12. Performance Improvement of Preemptive Discipline over Threshold Queueing for Dual Printers with 300 lps Secondary Printer

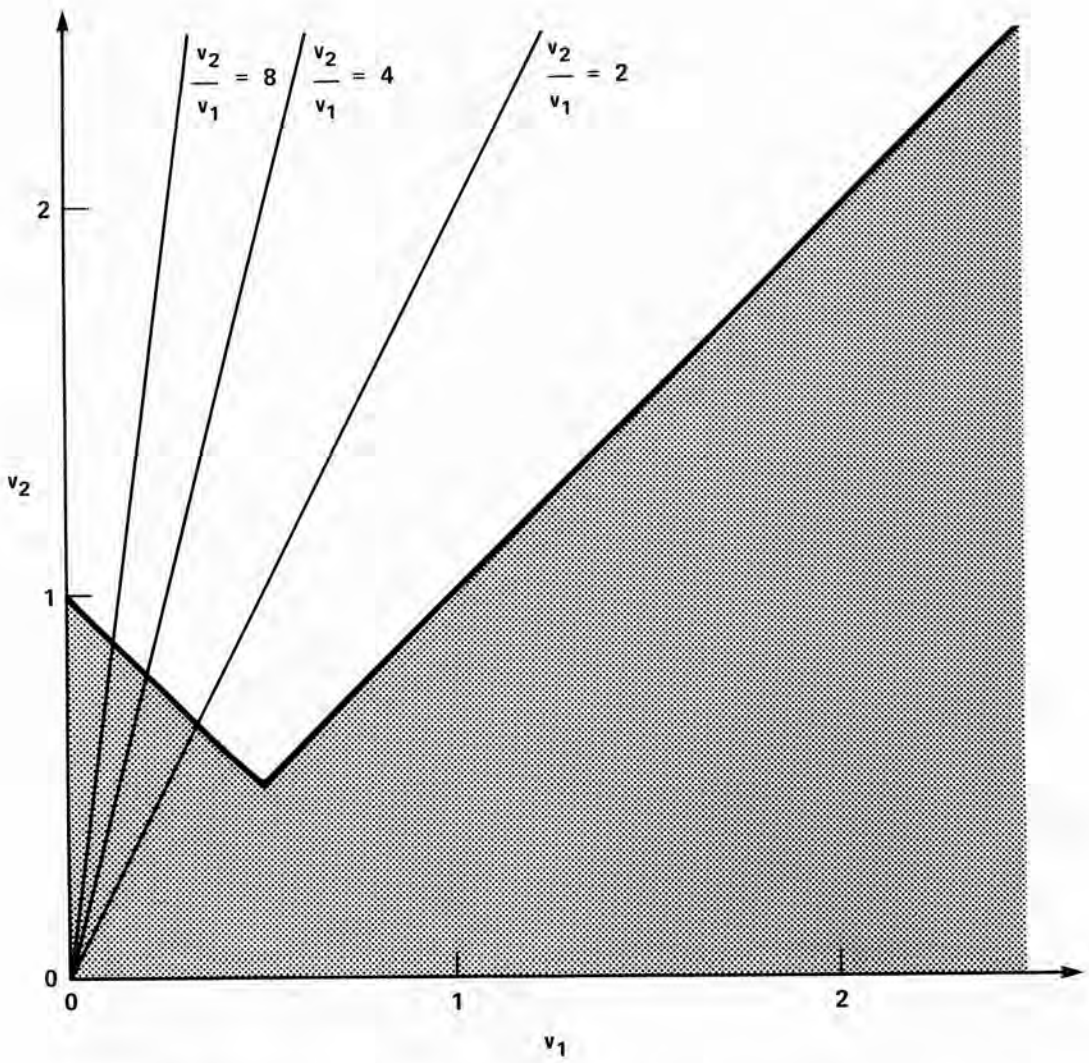



Figure 5.2-1. Region of $(v_1, v_2)$ plane for slow server preference

primary server for a system configuration in which the secondary server is four times as fast as the primary server. $\bar{N}$ is plotted as a function of $\lambda/\mu_1$ for threshold values of M from 1 to 8. By selectively varying M as a function of $\lambda/\mu_1$ one may adjust the mean performance of this system configuration along these curves. For an example, we consider the objective of maintaining $\bar{N}$ near 1. As shown in figure 5.2-3, we can construct a sawtooth composite curve with $\bar{N} = 1$ as the midpoint. The switchover points from one M-curve to the next (indicated by vertical arrows on the $v_1$ axis) identify $\lambda/\mu_1$ values at which the threshold value is adjusted to meet the stated objective. The cost of controlling performance

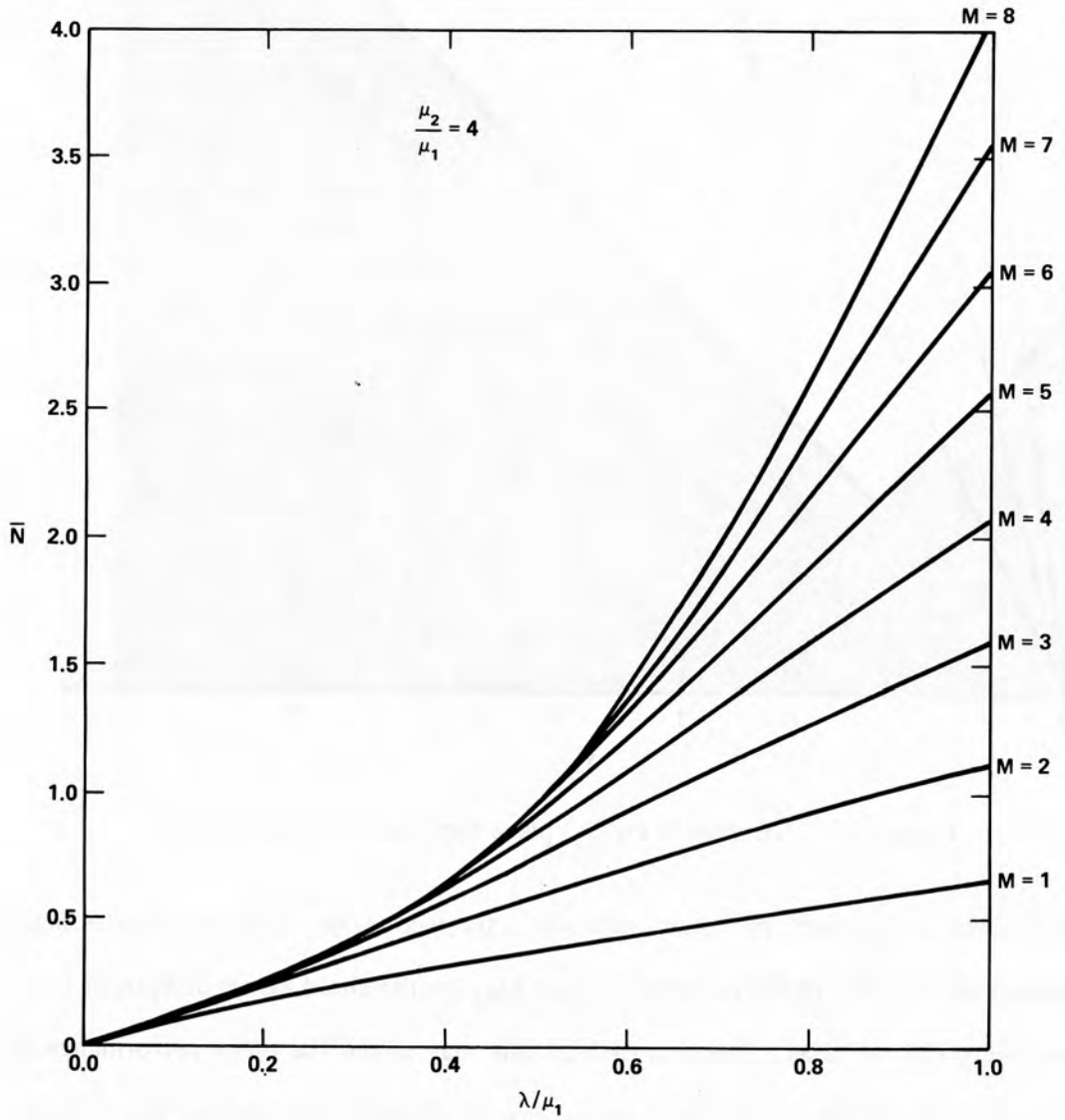



Figure 5.2–2. Performance of Secondary Server Preference with Constant Thresholds and $\mu_2/\mu_1 = 4$

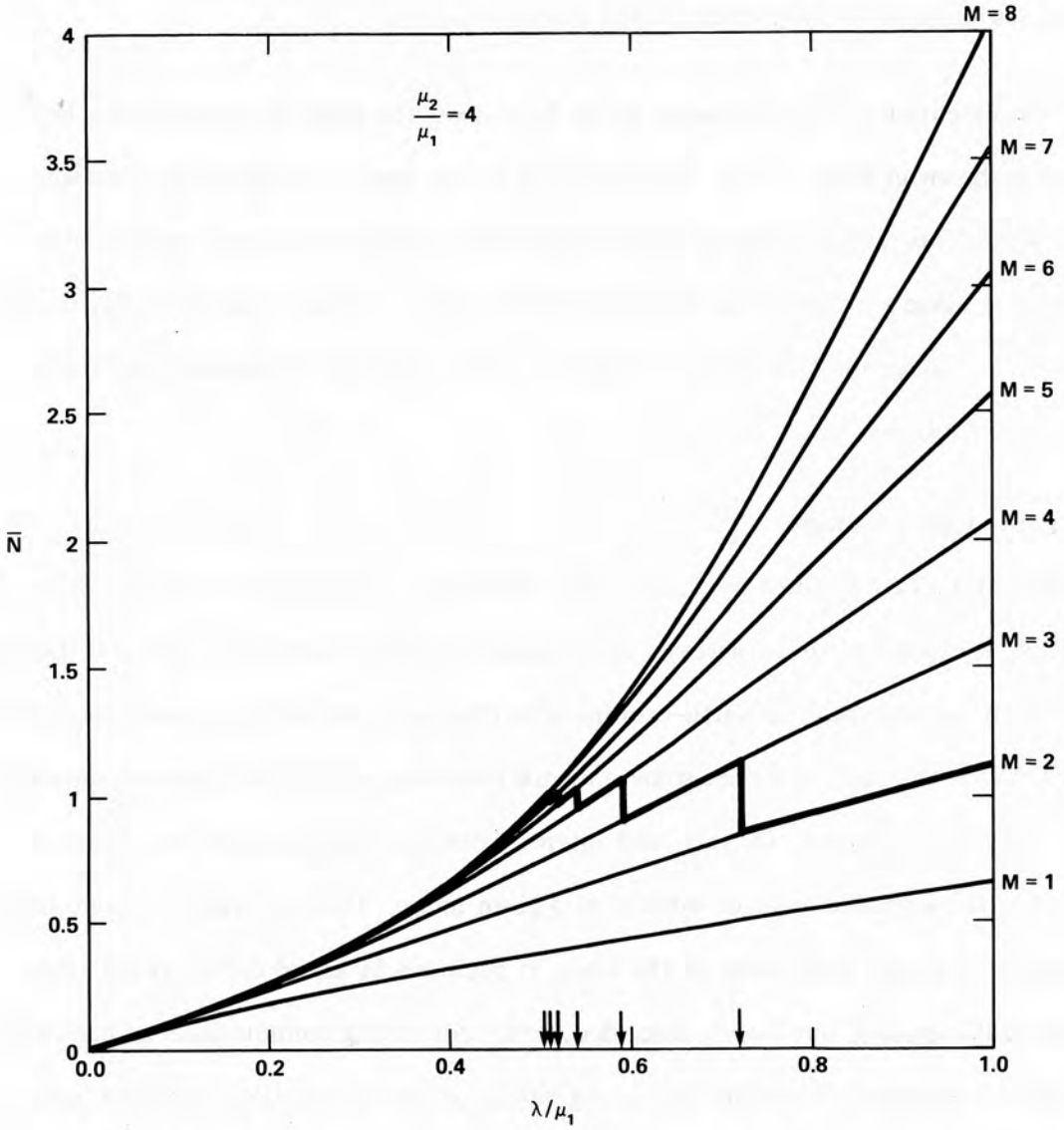



Figure 5.2–3. Regulating System Performance by Controlling the Threshold ($\lambda \leq \mu_1$)

in this manner is the utilization $\rho_2$ of the faster server. The corresponding utilization curves are shown in figure 5.2–4. Figure 5.2–5 superimposes the projection of the sawtooth curve onto the utilization curves to demonstrate that in such a configuration, relatively stable performance can be provided by a slow server depending only modestly on the resources of a faster server during loading transients.

The extension of $\bar{N}$ as shown in figure 5.2–3 over the range of operation for both servers is shown in figure 5.2–6. Superimposed on the figure is a sawtooth composite curve which illustrates a means of providing relatively stable performance over a wide operational range for the two server configuration. Such a strategy would be useful, for example, to reduce the load on a fast machine while providing a consistent and stable level of performance.

## 5.3 Finite Queue Capacity

In a system constrained by finite queue capacity, a faster server can be utilized as an overload relief mechanism to decrease the likelihood of queue overflow. The specific example to be considered is an earth–orbiting scientific spacecraft which acquires image data (e.g., NASA's Landsat), and does some onboard processing of the data prior to transmitting it to ground stations. Of particular interest is onboard information extraction to determine the scientific value or interest of a given image. Data of negligible value (e.g., its content is totally predictable or the image is occluded by cloud cover), is not transmitted to the ground, but merely discarded, hence conserving communications bandwidth and ground resources. The time consuming nature of the information extraction process coupled with the limited storage capacity onboard the spacecraft result in a possibility that the information extraction process may at times be unable to keep up with the flow of data into the spacecraft storage. To minimize the likelihood of losing potentially interesting data, a bypass mechanism is considered in which raw image data is transmitted directly to the ground without undergoing onboard processing. Of interest is determining

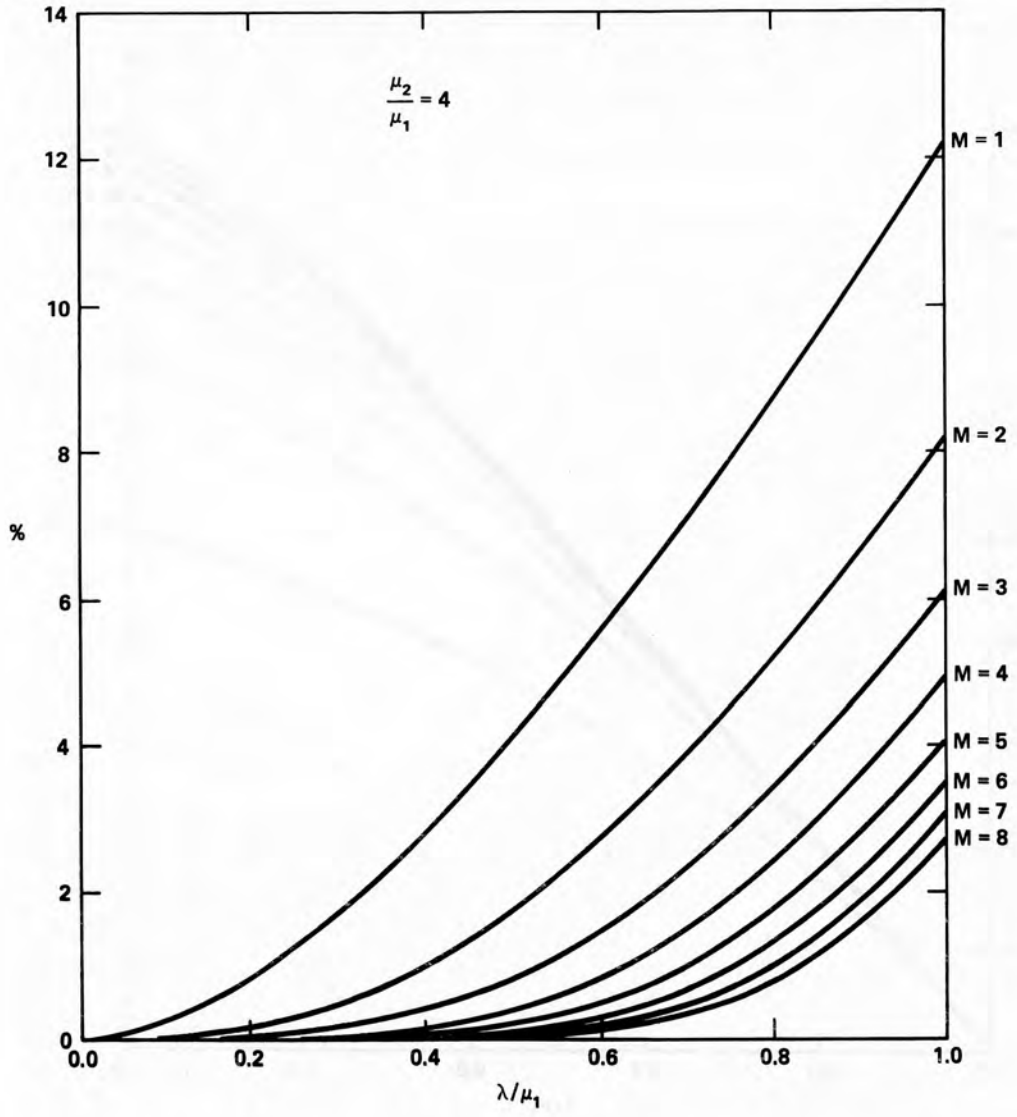



Figure 5.2–4(a). Utilization of Fast Server ($\rho_2$) under Slow Server Preference

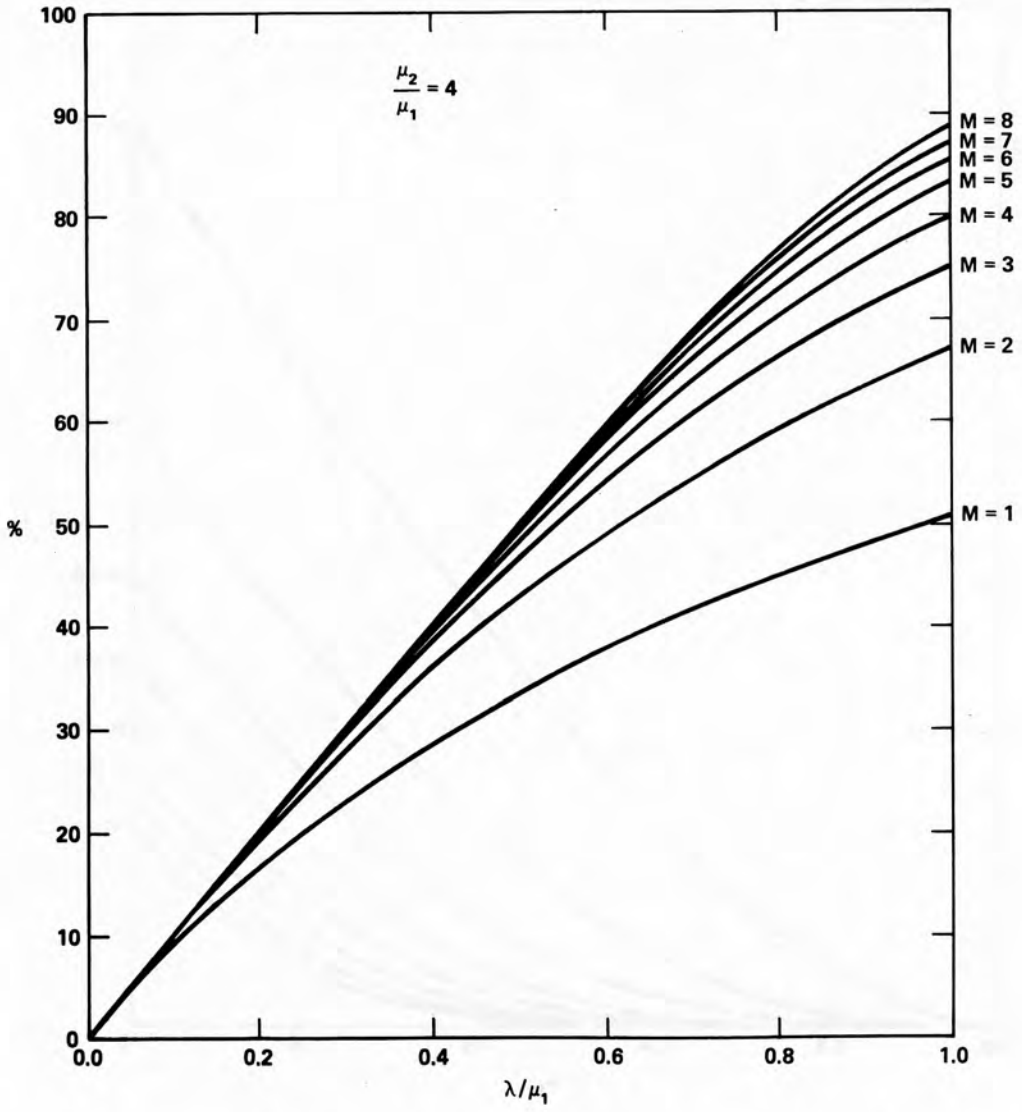



Figure 5.2–4(b). Utilization of Slow Server ($\rho_1$) under Slow Server Preference

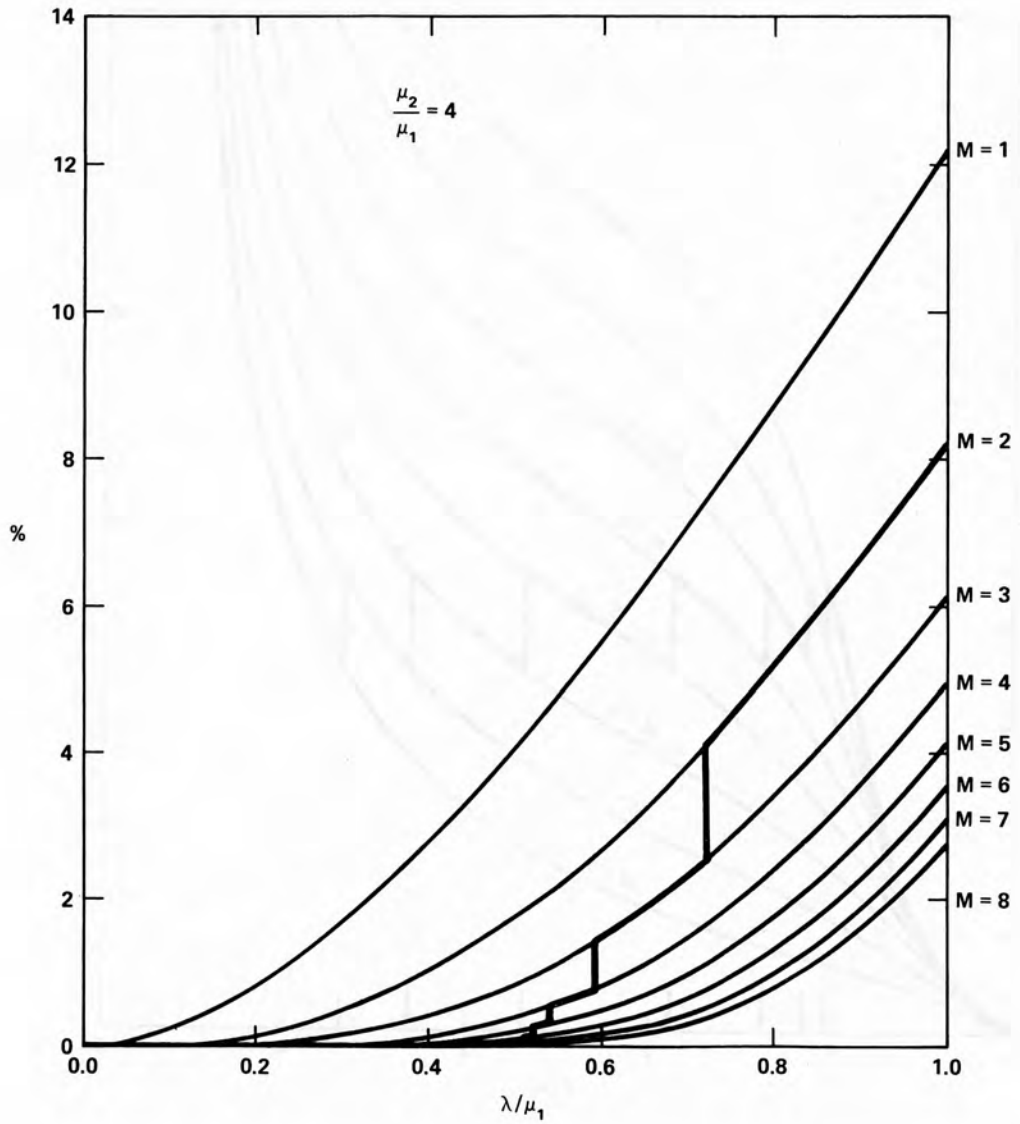



Figure 5.2-5. Fast Server Utilization ($\rho_2$) to Maintain $\overline{N}$ Near 1

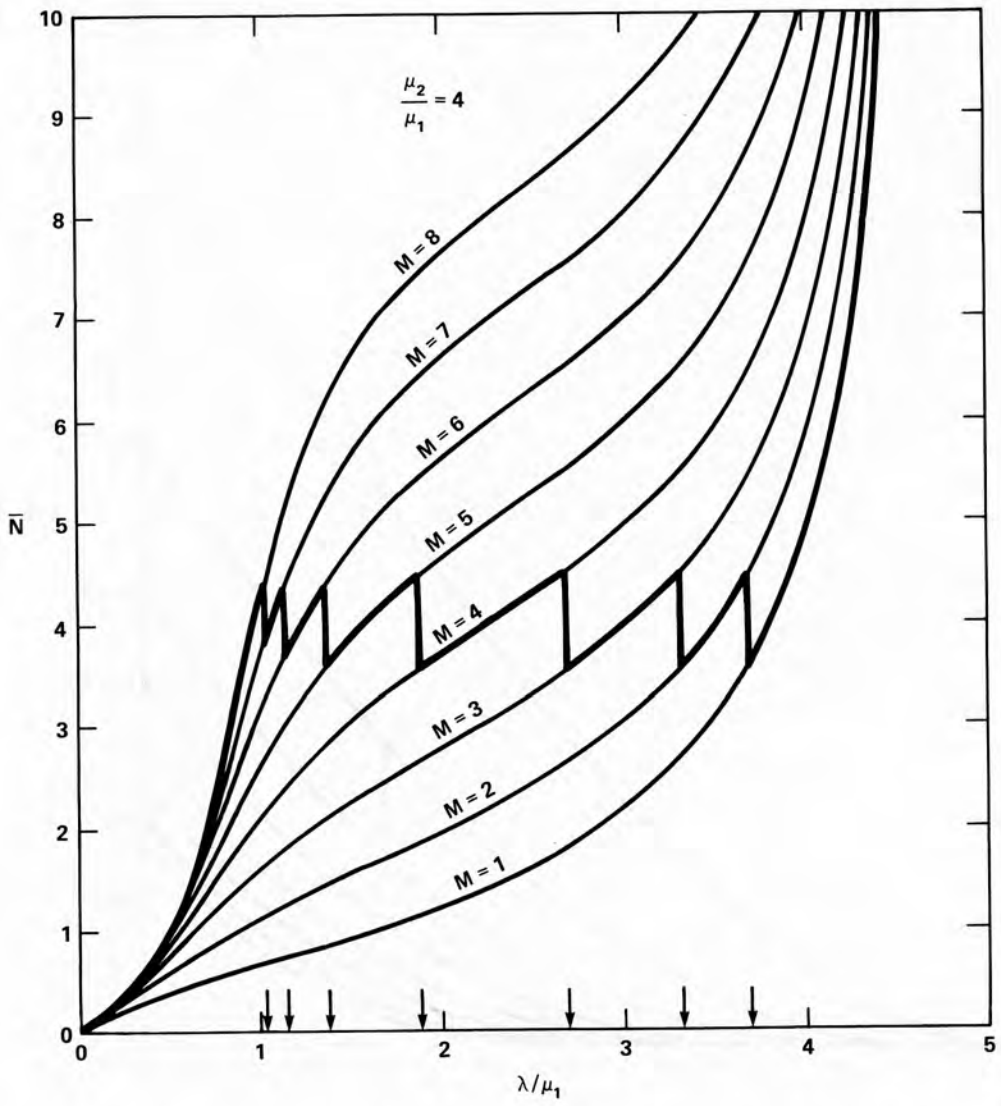



Figure 5.2–6. Regulating System Performance by Controlling the Threshold ($\lambda < \mu_1 + \mu_2$)

the criteria under which the bypass mechanism should be invoked. In this example, the threshold queueing formulation is extended to accommodate finite queues, and a typical spacecraft application is considered.

The results of chapter 3 can be readily extended to treat the existence of a finite rather than infinite capacity queue. As in the examples considered heretofore, deterministic policies will be considered. The state transition rate diagram is shown in figure 5.3–1. The infinite chain of states treated previously has been replaced with a truncated chain, the last state (Q, 11) signifying Q customers in the queue and both servers busy. Customers arriving while the system is in state (Q, 11) will be rejected. We are primarily concerned with the likelihood that customer rejection occurs, i.e., the probability that the system is in state (Q, 11) and a new customer arrives.

The steady state equations corresponding to figure 5.3–1 are:

a) $p_{0,00} = v_1 p_{0,10} + v_2 p_{0,01}$ (1)

b) $(1 + v_1)p_{0,10} = p_{0,00} + v_1 p_{1,10} + v_2 p_{0,11}$

c) $(1 + v_1)p_{i,10} = p_{i-1,10} + v_1 p_{i+1,10} + v_2 p_{i,11} \quad 1 \leq i \leq M - 2$

d) $(1 + v_1)p_{M-1,10} = p_{M-2,10} + v_2 p_{M-1,11}$

e) $(1 + v_2)p_{0,01} = v_1 p_{0,11}$

f) $(1 + v_1 + v_2)p_{0,11} = p_{0,01} + v_1 p_{1,11}$

g) $(1 + v_1 + v_2)p_{i,11} = p_{i-1,11} + v_1 p_{i+1,11} \quad 1 \leq i \leq M - 2$

h) $(1 + v_1 + v_2)p_{M-1,11} = p_{M-2,11} + p_{M-1,10} + (v_1 + v_2)p_{M,11}$

i) $(1 + v_1 + v_2)p_{M+i,11} = p_{M+i-1,11} + (v_1 + v_2)p_{M+i+1,11} \quad 0 \leq i \leq Q - M - 1$

j) $p_{Q-1,11} = (v_1 + v_2)p_{Q,11}$

The solution to eqns (1) is:

a) $p_{i,11} = (f_{i+1} - f_i)p_{0,01} \quad 0 \leq i \leq M - 1$ (2)

b) $p_{0,00} = v_2 \sum_{i=0}^{M} v_1^i f_i \, p_{0,01}$

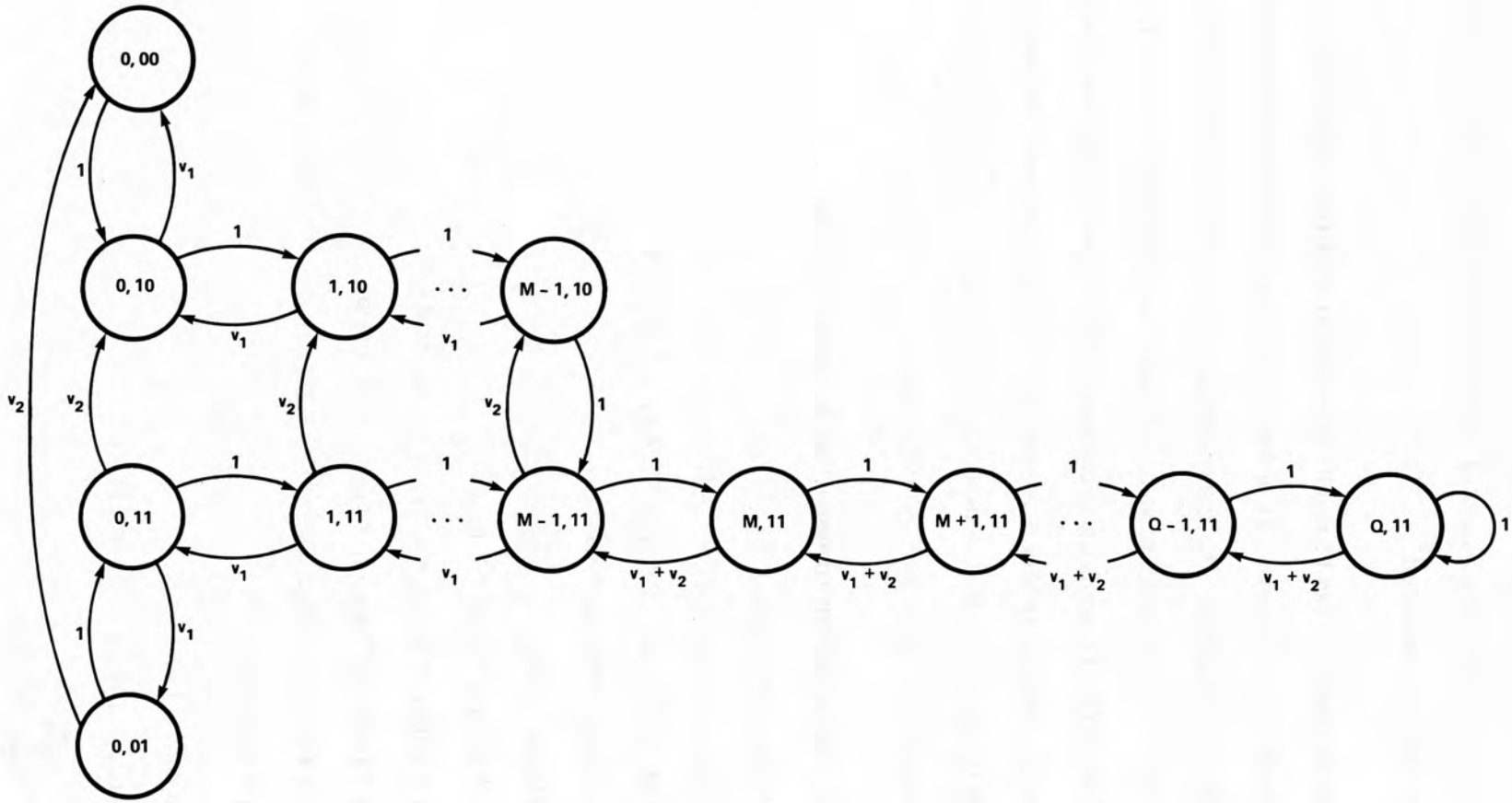



Figure 5.3–1. Threshold Queueing with Queue Capacity = Q

c) $$p_{i,10} = v_2 \sum_{j=i+1}^{M} v_1^{j-i-1} f_j p_{0,01} \qquad 0 \le i \le M-1$$

d) $$p_{M+i,11} = \left(\frac{1}{v_2 + v_1}\right)^{i+1} (f_M - f_{M-1}) p_{0,01} \qquad 0 \le i \le Q-M$$

e) $$p_{0,01}^{-1} = v_2 \sum_{i=0}^{M} v_1^i f_i + v_2 \sum_{i=1}^{M} \sum_{j=0}^{i-1} v_1^j f_i + f_M + \sum_{i=0}^{Q-M} \left(\frac{1}{v_2 + v_1}\right)^{i+1} (f_M - f_{M-1})$$

f) $$f_i = \sum_{j=0}^{\lfloor \frac{i}{2} \rfloor} (-1)^j \binom{i-j}{j} \left(\frac{v_2 + v_1 + 1}{v_1}\right)^{i-2j} v_1^{-j}$$

The probability of customer rejection is the probability that the system is in state (Q, 11) when a customer arrives:

$$\begin{aligned} P[\text{customer rejection}] = P[\text{customer arrives} \mid \text{system is in state } (Q, 11)] \times \\ P[\text{system is in state } (Q, 11)] \end{aligned} \tag{3}$$

In state (Q, 11), the system will move to state (Q - 1, 11) before the next customer arrival with probability $\frac{v_2 + v_1}{v_2 + v_1 + 1}$. With probability $\frac{1}{v_2 + v_1 + 1}$, a new customer will arrive before the transition to (Q - 1, 11) takes place, so we have:

$$P[\text{customer rejection}] = \left(\frac{1}{v_2 + v_1 + 1}\right) p_{Q,11}$$

$$= \left(\frac{1}{v_2 + v_1 + 1}\right)\left(\frac{1}{v_2 + v_1}\right)^{Q-M+1} (f_M - f_{M-1}) \left[ v_2 \sum_{i=0}^{M} v_1^i f_i + v_2 \sum_{i=1}^{M} \sum_{j=0}^{i-1} v_1^j f_i + f_M + \sum_{i=0}^{Q-M} \left(\frac{1}{v_2 + v_1}\right)^{i+1} (f_M - f_{M-1}) \right]^{-1} \tag{4}$$

The mean queue length for a finite queue system with queue capacity Q is:

$$\bar{Q}(Q, M, v_1, v_2) = \sum_{i=1}^{M-1} i p_{i,10} + \sum_{i=1}^{Q} i p_{i,11}$$

$$= \frac{v_2 \sum_{i=1}^{M-1} i \sum_{j=i}^{M-1} v_1^{j-i} f_j + M f_M - \sum_{i=1}^{M} f_i + (f_M - f_{M-1}) \sum_{i=0}^{Q-M} (M + i)(v_2 + v_1)^{-i-1}}{v_2 \sum_{i=1}^{M} \sum_{j=0}^{i-1} v_1^j f_i + f_M + v_2 \sum_{i=0}^{M} v_1^i f_i + (f_M - f_{M-1}) \sum_{i=0}^{Q-M} (v_2 + v_1)^{-i-1}} \tag{5}$$

As described previously, we consider a spacecraft data management application to illustrate the utility of threshold queueing to manage a finite queue system. As onboard sensor data processing is a relatively new concept, which has yet to be attempted, no actual date is available on which to base a rigorous analysis. The example to be considered is a hypothetical one, built on parameters typical of Landsat spacecraft and operational experience from the Landsat missions.

We consider a spacecraft with seven imaging sensors, in which one scene consists of a 6000 X 6000 pixel array (one byte/pixel) from each sensor. One composite scene is then 252 Mbytes of data (36 Mbytes from each of seven sensors). A typical onboard algorithm of interest is a cloud cover algorithm, which discards images which have cloud cover in excess of a predefined level. This algorithm requires only one sensor's input, and typically requires a couple instructions per pixel to process the image, or about 72M instructions. To keep the example simple, we hypothesize the existence of a small computer onboard the spacecraft which operates at an instruction rate of about 720K instructions per second, so that the processing of one image requires typically 100 seconds. We further hypothesize the existence of a 10 Mbyte/second transmitter which can remove raw images (252 Mbytes) from the queue and transmit them to earth in about 25 seconds. This combination of processor (primary, slow server) and transmitter (secondary, fast server) results in a server ratio $\mu_2/\mu_1 = 4$. Image buffer sizes of 4 and 8 images are considered.

In figure 5.3–2, the probability of buffer overflow is plotted as a function of normalized arrival rate $\lambda/\mu_1$ for the case of no auxiliary transmitter, and for the case where there is such a transmitter. The buffer capacity for this figure is 4, and the four curves reflecting use of the auxiliary transmitter depict the overflow probability for each of the four possible threshold sizes. Figure 5.3–3 shows the proportion of the input images which are actually processed onboard (primary server) versus those that are sent to the ground unprocessed (secondary server). Figures 5.3–4 and 5.3–5 depict the same system

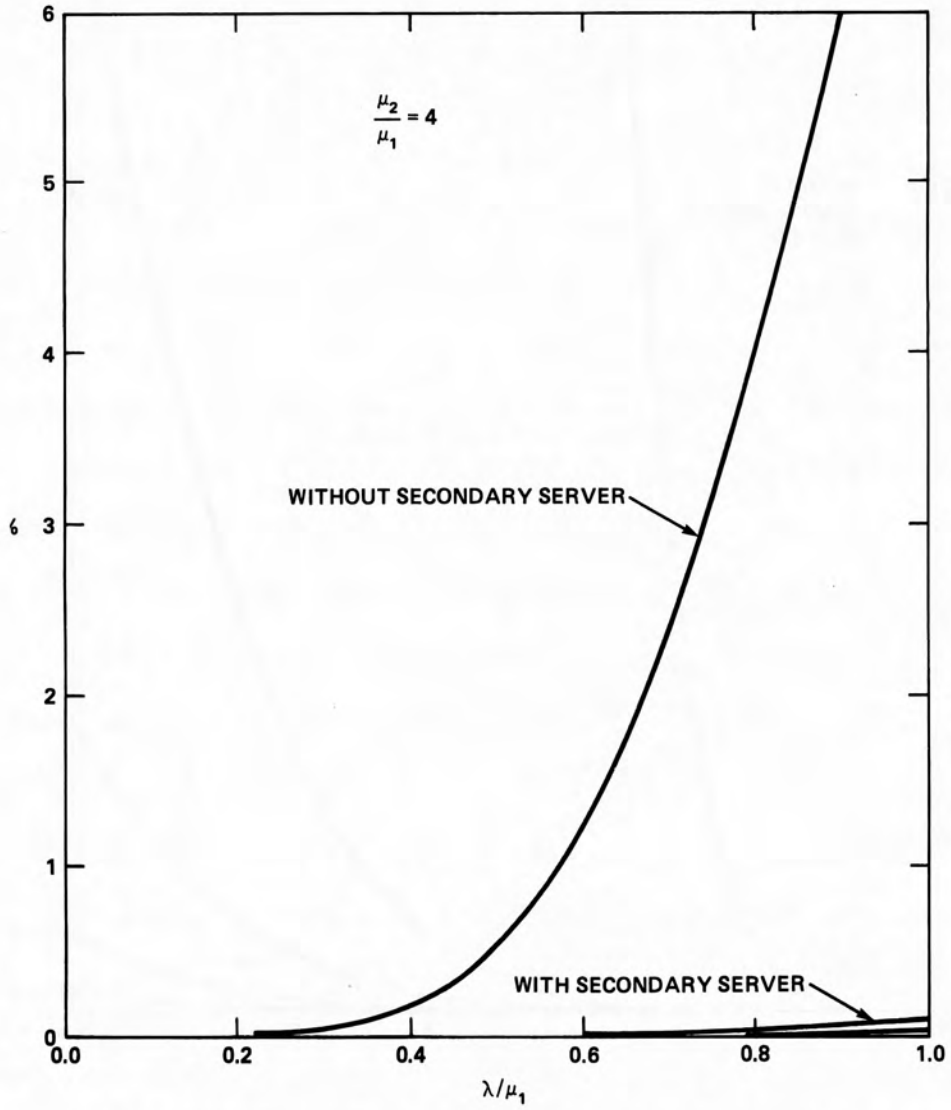



Figure 5.3-2(a). Probability of Queue Overflow for Queue Capacity of 4 with $\mu_2/\mu_1 = 4$

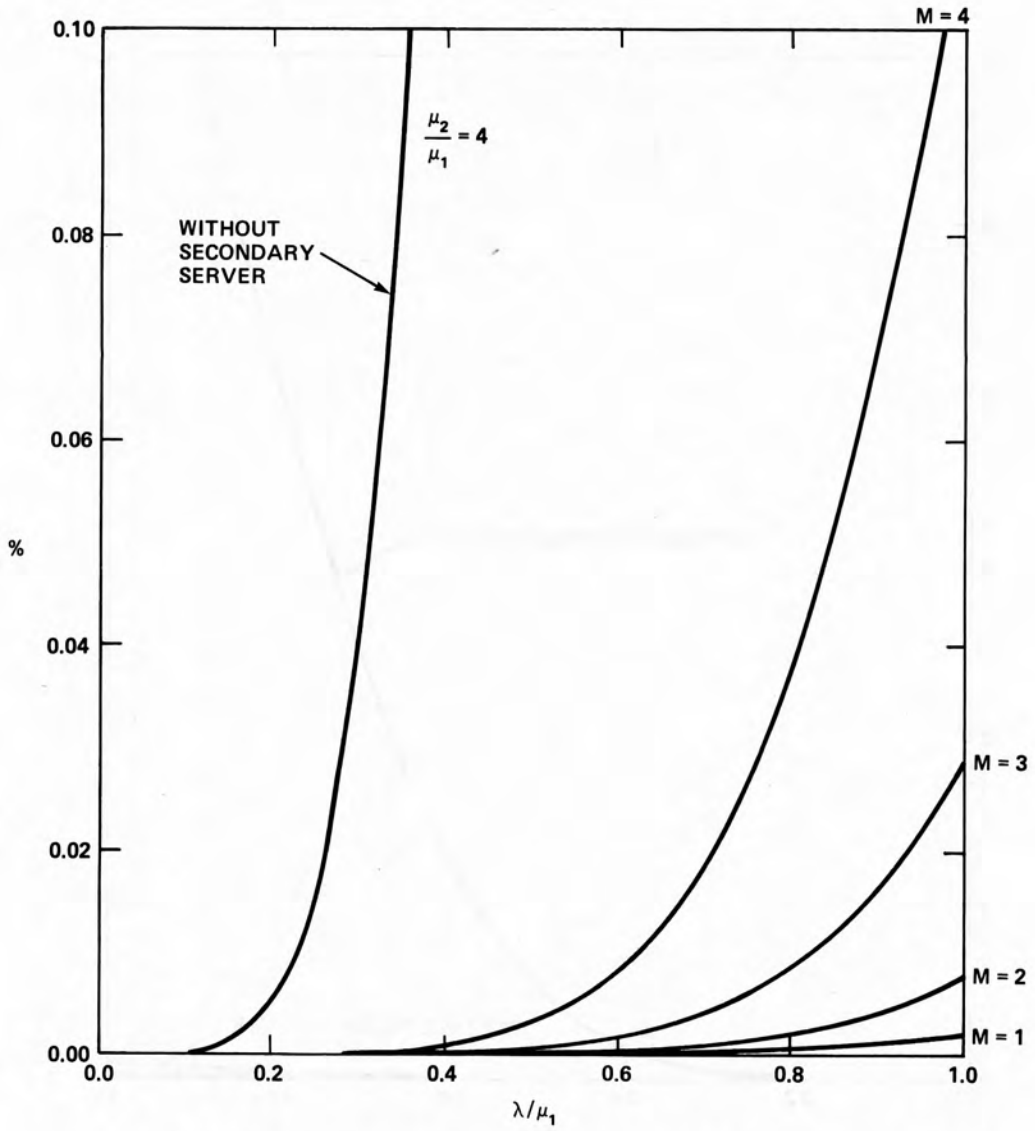



Figure 5.3-2(b). Probability of Queue Overflow for Queue Capacity of 4 with $\mu_2/\mu_1 = 4$ (Detail)

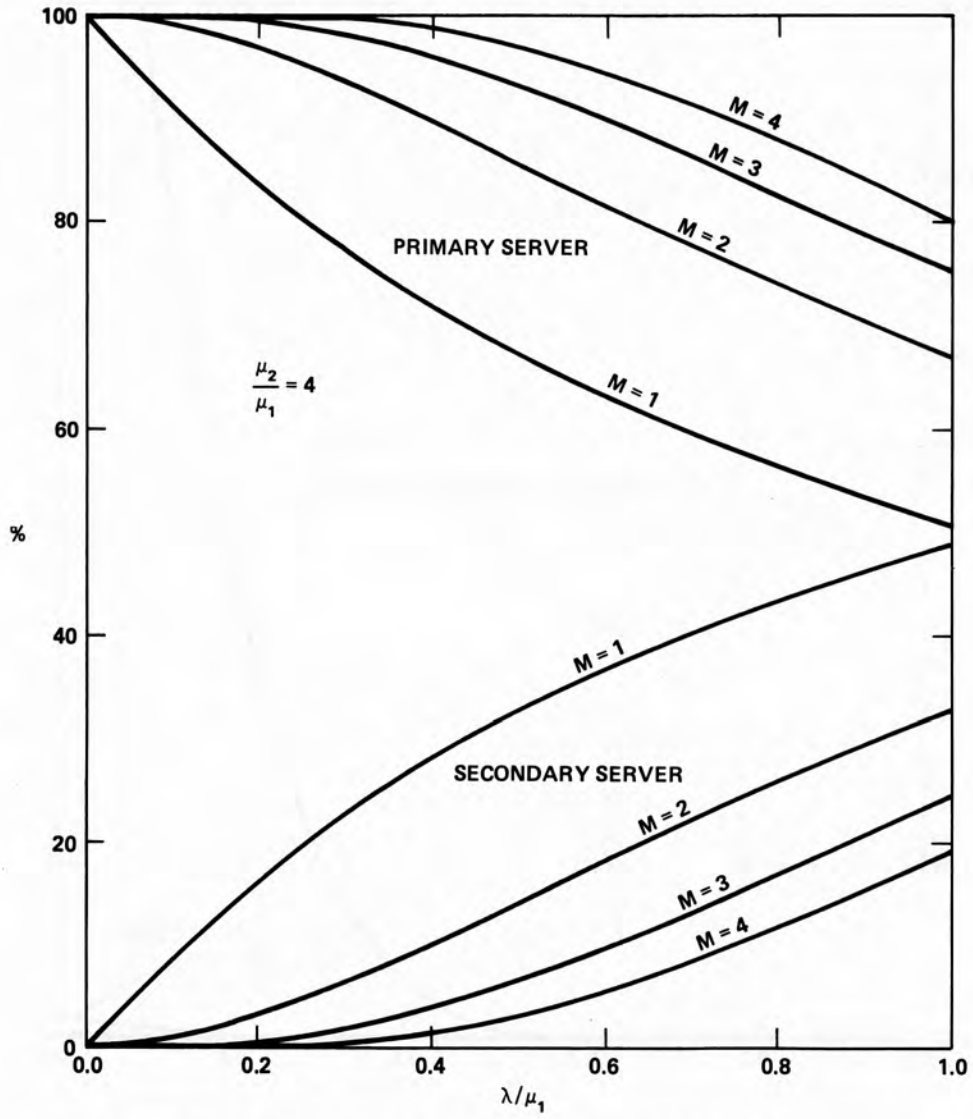



Figure 5.3–3. Distribution of Work Flow through Primary and Secondary Servers for $\mu_2/\mu_1 = 4$ with Queue Capacity of 4

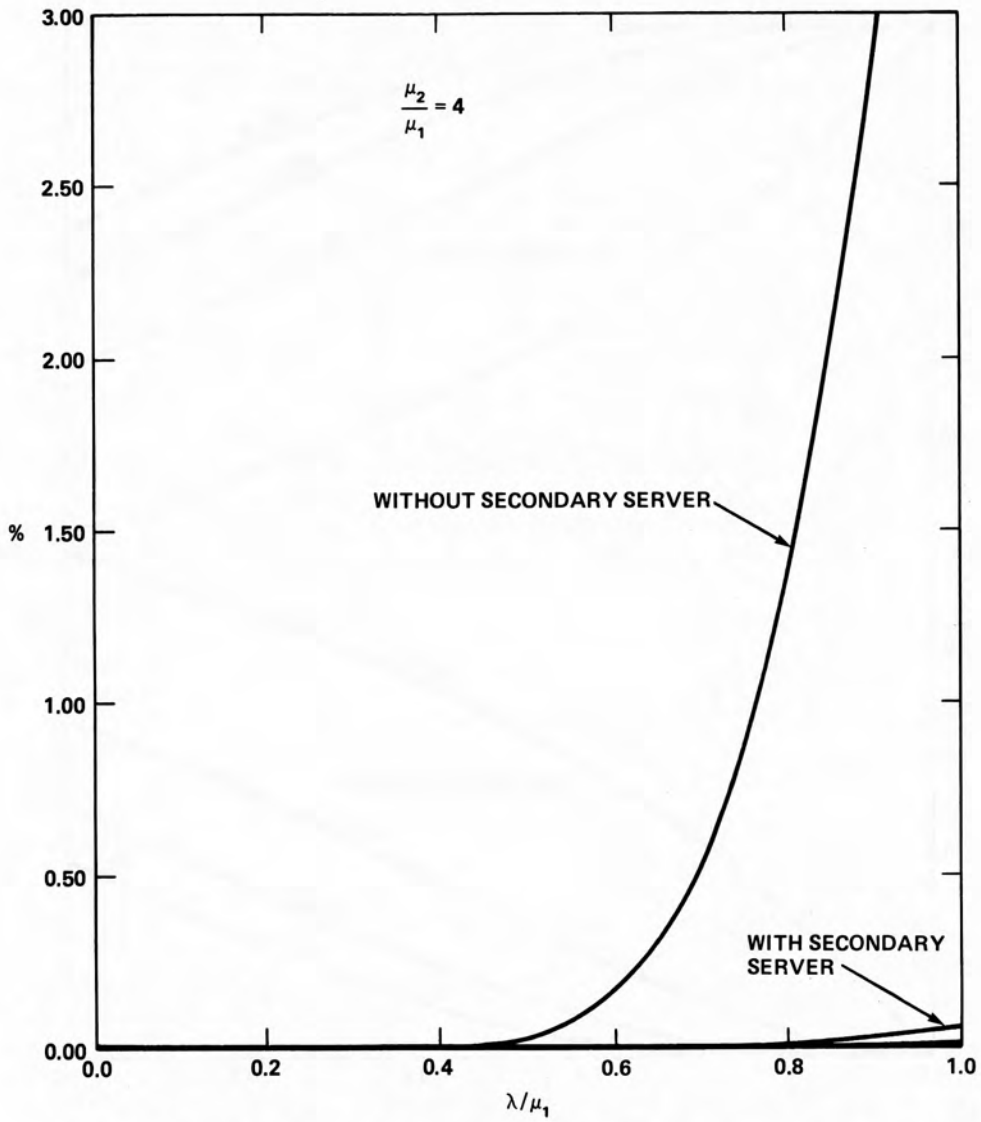



Figure 5.3–4(a). Probability of Queue Overflow for Queue Capacity of 8 with $\mu_2/\mu_1 = 4$

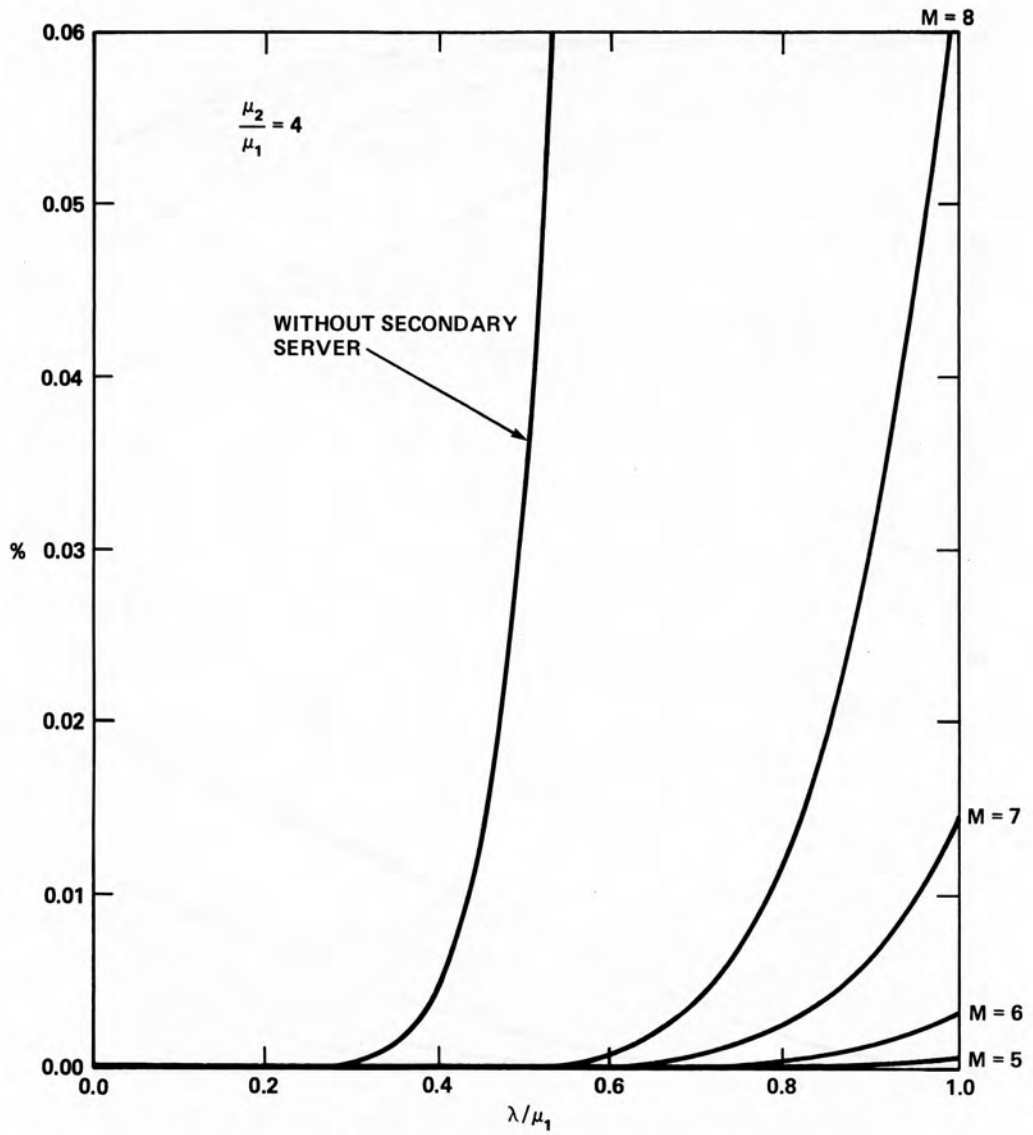



Figure 5.3–4(b). Probability of Queue Overflow for Queue Capacity of 8 with $\mu_2/\mu_1 = 4$ (Detail)

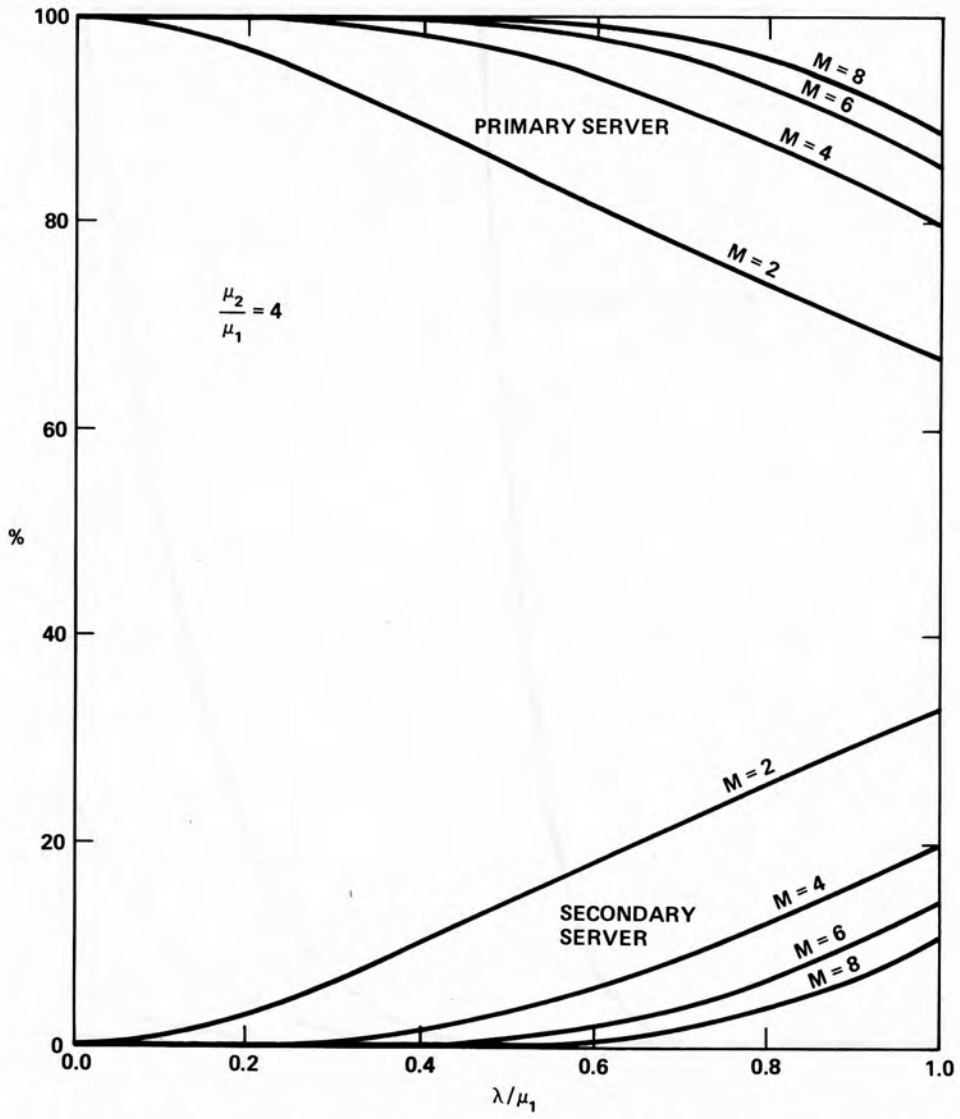



Figure 5.3-5. Distribution of Work Flow through Primary and Secondary Servers for $\mu_2/\mu_1 = 4$ with Queue Capacity of 8

but with the buffer storage enlarged to accommodate 8 images. Looking more closely at this case, we see that with a threshold of M = 6, less than .01% of the input images will be lost while nearly 90% of them can be processed onboard over the range of operation of the system. We see, therefore, that threshold queueing provides an effective control discipline for managing a system with limited buffer storage.

## 5.4 Threshold Queueing for n-Servers

The difficulties involved in extending the state transition rate diagram approach to the n-server problem become apparent upon constructing the sequence of state diagrams successively for n = 2, 3, and 4. These are shown in figures 5.4-1, 2, and 3. The combinatoric growth is such that for each successive value of n, the complete problem must be solved from the beginning, with the complexity growing proportional to $2^n$. Rather than taking this approach, we will present a more general approach for analyzing a specific system configuration for arbitrary n.

We recall from queueing theory [KLEIL75] that one can construct a transition probability matrix T and a state probability vector P which together totally specify the operation of the system. If $P^{(0)}$ is the initial probability associated with each system state, then

$$P^{(1)} = P^{(0)}T \tag{1}$$

is the system state probabilities after one state transition, and

$$P^{(i)} = P^{(i-1)}T = P^{(0)}T^i \tag{2}$$

gives the system state probabilities after i state transitions. The steady state probabilities associated with each state are then:

$$P^{(\infty)} = \lim_{i \to \infty} P^{(i)} = P^{(0)} \lim_{i \to \infty} T^i \tag{3}$$

These probabilities are also the solution to the equation

$$P = PT \tag{4}$$

To utilize this in the n-server threshold queueing problem, we first must perform a mapping from the current state designation to the positive integers, then convert the continuous

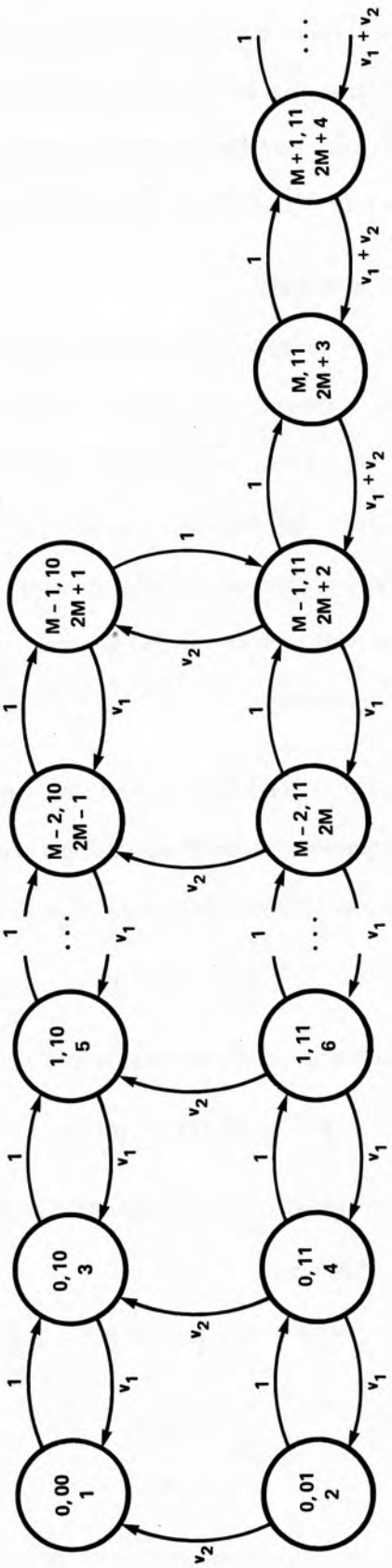



Figure 5.4–1. State Transition Rate Diagram for Two Servers

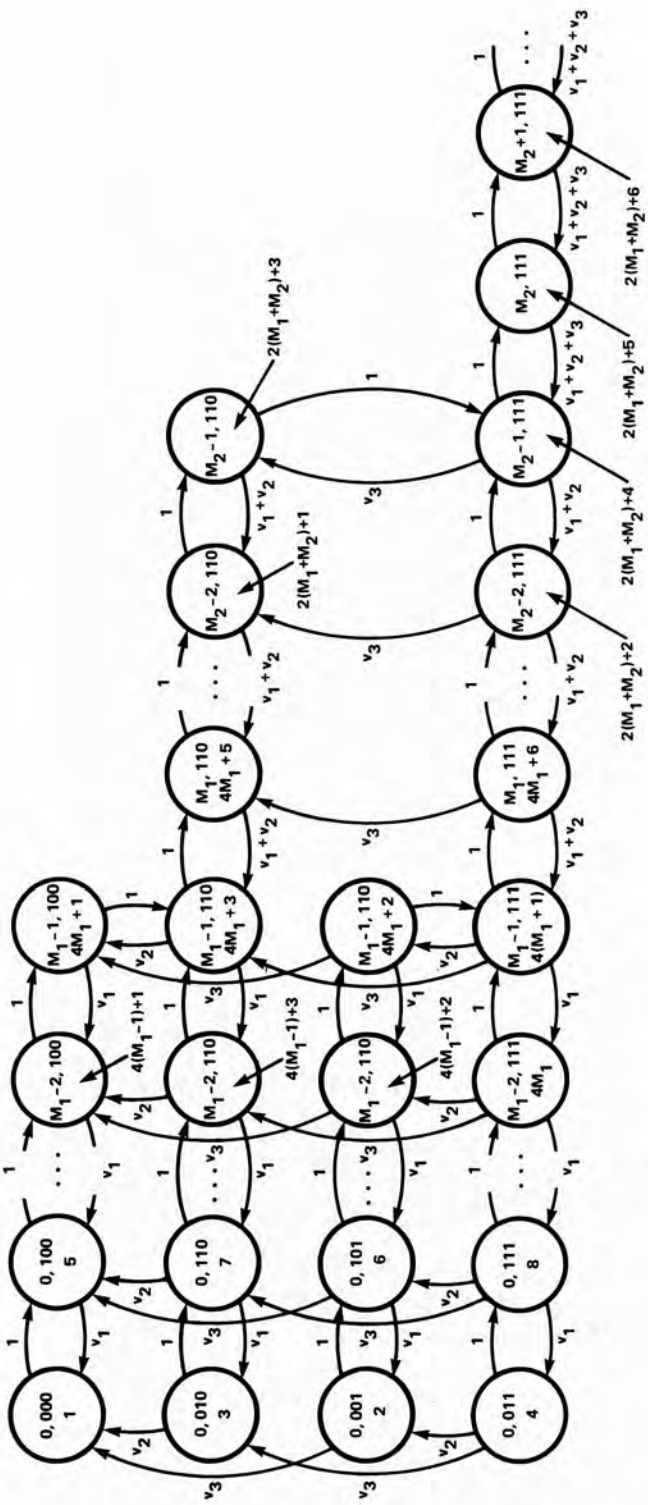



Figure 5.4–2. State Transition Rate Diagram for Three Servers

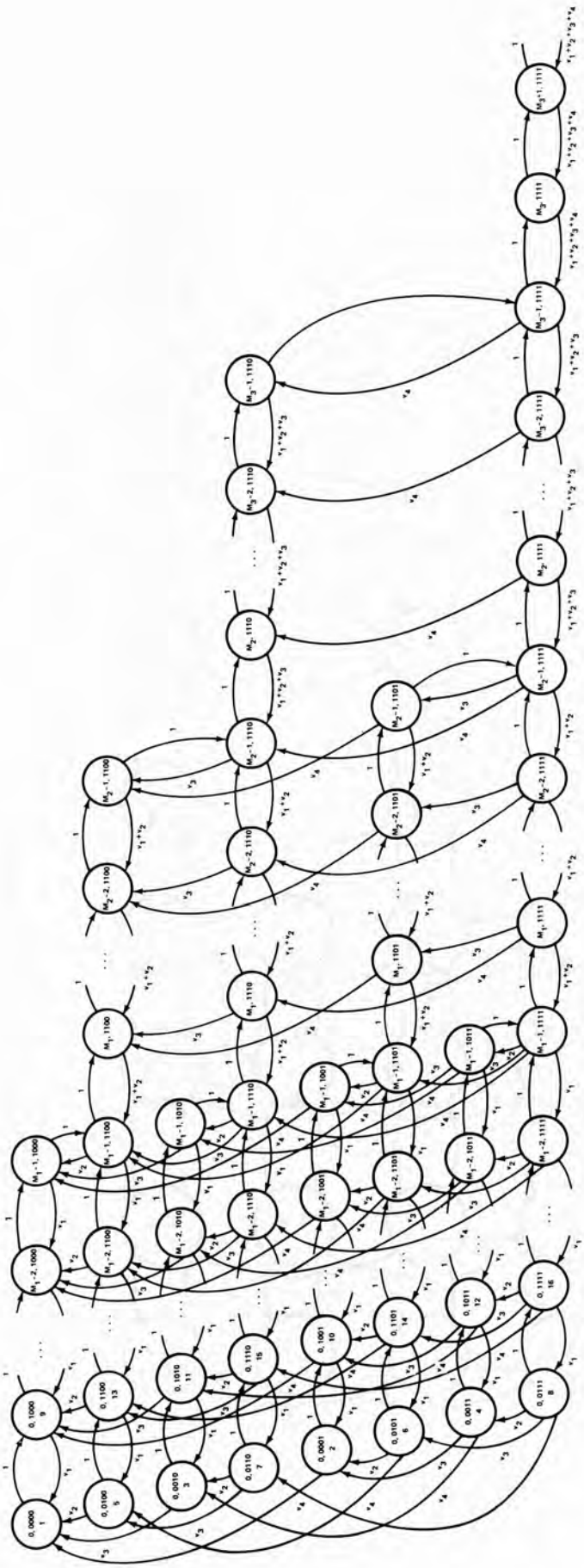
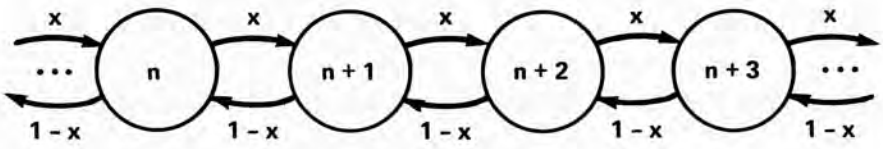


Figure 5.4–3. State Transition Rate Diagram for Four Servers

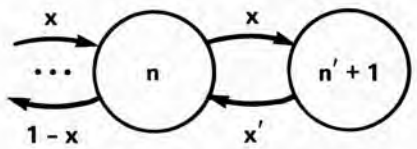
time model with transition rates into its equivalent discrete time model with transition probabilities [SERFR79], and finally convert the infinite state model to an equivalent finite state model. The state numbering scheme to be employed is shown in figures 5.4–1, 2, and 3 in conjunction with the state identification utilized previously. The conversion to the equivalent discrete time model is accomplished by dividing all transition rates in the continuous time model by $1 + \sum_{i=1}^{n} v_i$ and setting

$$t_{ii} = 1 - \sum_{\substack{j=1 \\ j \neq i}}^{n} t_{ij} \tag{5}$$

Now to truncate the infinite state model to yield a finite state model, we note that the necessary step is to convert an infinite state transition probability diagram of the form




into a finite state transition probability diagram of the form




which yields identical steady state probabilities for all states up to and including n.* To do this, we note that the steady state probability associated with state n + i is:

$$p_{n+i} = \left(\frac{x}{1-x}\right)^{i} p_n \qquad i \geq 0 \tag{6}$$

and

$$p_{n'+1} = \sum_{i=1}^{\infty} p_{n+i} = p_n \sum_{i=1}^{\infty} \left(\frac{x}{1-x}\right)^{i} = p_n \left(\frac{x}{1-2x}\right) \tag{7}$$

*Computation of performance parameters such as $\bar{N}$ requires consideration of the full Markov chain, as discussed in section 3.3 and Appendix C.

where it has been assumed that $x < \frac{1}{2}$, which is the ergodic assumption. Now we can compute x′ directly:

$$x' = \frac{p_{n+1}}{p_{n'+1}} (1 - x) = 1 - 2x \tag{8}$$

Performing these three steps on the n-server state transition rate diagram to produce the transition probability matrix T yields the following:

$$T = \left(1 + \sum_{i=1}^{n} v_i\right)^{-1} (Q + D) \tag{9}$$

where the elements of D are:

$$d_{ij} = \begin{cases} 0 \text{ for } i \neq j \\ 1 + \sum_{k=1}^{n} v_k - \sum_{k=1}^{N} q_{ik} \text{ for } i = j, \text{ where n = the no. of servers and N = the no. of columns in Q} \end{cases} \tag{10}$$

The Q matrix is illustrated on the following page, utilizing the following notation:

$$E_i^n = \begin{bmatrix} [E_{i+1}^{n-1}][O^{n-1}] \\ [v_i I^{n-1}][E_{i+1}^{n-1}] \end{bmatrix}$$

$$E_{i+n}^0 = [O^0]$$

$I^n = 2^n \times 2^n$ identity matrix

$O^n = 2^n \times 2^n$ matrix of zeroes

$$\hat{I}^n = \begin{bmatrix} [O^{n-1}] & [I^{n-1}] \\ [O^{n-1}] & [O^{n-1}] \end{bmatrix}$$

$$\check{I}^n = \begin{bmatrix} [O^{n-1}] & [O^{n-1}] \\ [I^{n-1}] & [O^{n-1}] \end{bmatrix}$$

$$
Q = \begin{bmatrix}
\left[ E_2^{n-1} \right] \left[ I^{n-1} \right] & & & & & & \\
\left[ v_1 I^{n-1} \right] \left[ E_2^{n-1} \right] \left[ I^{n-1} \right] & & & & & & \\
\left[ v_1 I^{n-1} \right] \left[ E_2^{n-1} \right] \left[ I^{n-1} \right] & & & & & & \\
\vdots & & & & & & \\
\left[ v_1 I^{n-1} \right] \left[ E_2^{n-1} \right] \left[ I^{n-1} \right] & & & & & & \\
\left[ v_1 I^{n-1} \right] \left[ E_2^{n-1} + \hat{I}^{n-1} \right] \left[ \check{I}^{n-1} \right] & & & & & & \\
\left[ O^{n-2} \right] \left[ (v_2 + v_1) I^{n-2} \right] \left[ E_3^{n-2} \right] \left[ I^{n-2} \right] & & & & & & \\
\left[ (v_2 + v_1) I^{n-2} \right] \left[ E_3^{n-2} \right] \left[ I^{n-2} \right] & & & & & & \\
\vdots & & & & & & \\
\left[ (v_2 + v_1) I^{n-2} \right] \left[ E_3^{n-2} \right] \left[ I^{n-2} \right] & & & & & & \\
\left[ (v_2 + v_1) I^{n-2} \right] \left[ E_3^{n-2} + \hat{I}^{n-2} \right] \left[ \check{I}^{n-2} \right] & & & & & & \\
\left[ O^{n-3} \right] \left[ (v_3 + v_2 + v_1) I^{n-3} \right] \left[ E_4^{n-3} \right] \left[ I^{n-3} \right] & & & & & & \\
\vdots & & & & & & \\
\left[ \sum_{i=1}^{n-1} v_i I^1 \right] \left[ E_n^1 \right] \left[ I^1 \right] & & & & & & \\
\left[ \sum_{i=1}^{n-1} v_i I^1 \right] \left[ E_n^1 + \hat{I}^1 \right] \left[ \check{I}^1 \right] & & & & & & \\
\left[ O^0 \right] \left[ \sum_{i=1}^{n} v_i I^0 \right] \left[ O^0 \right] \left[ I^0 \right] & & & & & & \\
\left[ \sum_{i=1}^{n} v_i - 1 \right] \left[ O^0 \right] & & & & & &
\end{bmatrix}
$$

$M_1 - 1$ rows

$M_1$ threshold

$M_2 - M_1 - 1$ rows

$M_2$ threshold

$M_{n-1}$ threshold

0

0

(11)

The dimension of Q is $M_{n-1} + 2 + 2^{n-1} + \sum_{i=1}^{n-1} 2^{i-1} M_{n-i}$ where $M_i$ is the $i^{th}$ threshold.

Now, given the number of servers n, the parameters $v_1, v_2, \ldots, v_n$, and the threshold sizes $M_1, M_2, \ldots, M_{n-1}$, the T matrix can be constructed, and the steady state probability vector P computed either iteratively by eqn (3) or directly via eqn (4). Given P, we can then easily compute $\bar{N}$ (the mean number of customers in the system), $\bar{Q}$ (the mean queue length), and $\rho_i$ (the utilization of server i).

Identification of optimal threshold values for the n-server system can be accomplished by a straightforward search of the space of feasible threshold values, through construction of the corresponding T matrices and solution for $\bar{N}$. The feasible space can readily be bounded, thereby bounding the resultant search process. To establish bounds on the $M_i$, the n-server system shown in figure 5.4–4 will be considered, where $\mu_1 \geq \mu_2 \geq \ldots \geq \mu_n$.

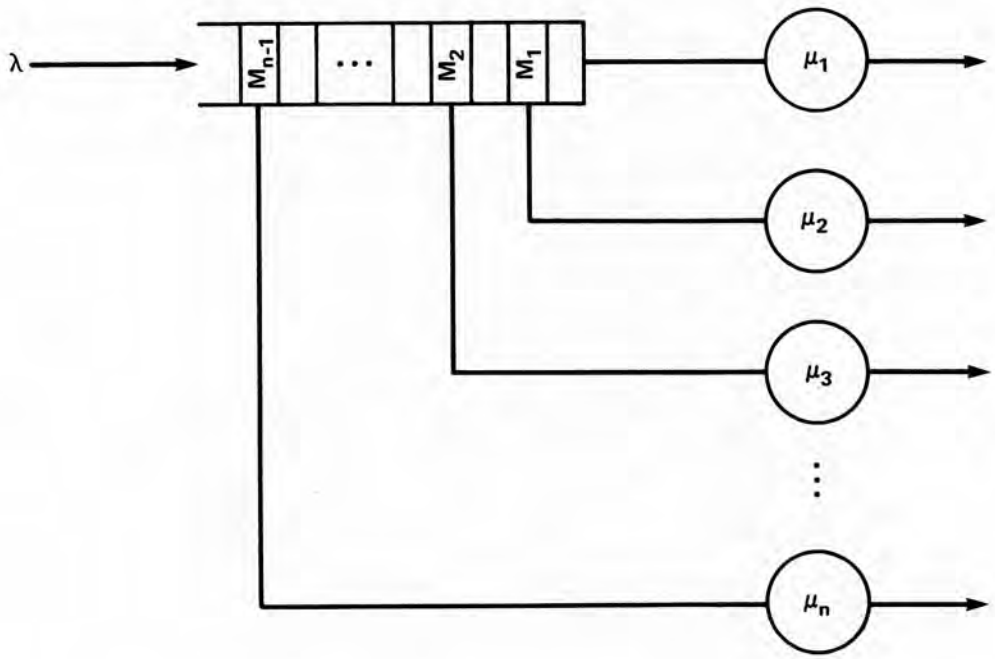



Figure 5.4–4. Multiple server threshold queueing configuration

We utilize the 2-server approximation to the threshold size given by eqn (4.2–50):

$$\widetilde{M} = \left\lfloor \frac{v_1 - 1 + \sqrt{4v_2 + (v_1 - 1)^2}}{2v_2} \right\rfloor \tag{12}$$

noting that in the limit as $v_1$ and $v_2 \to \infty$ along the line $v_1 = cv_2$, $\widetilde{M}_{max} \to \lfloor c \rfloor$. This limit corresponds to $\lambda \to 0$, i.e., very light loading. An upper limit on $M_1$ can be established simply by considering the two server configuration with $\mu_1$ and $\mu_2$:

$$M_1 \leq \left\lfloor \frac{\mu_1}{\mu_2} \right\rfloor \tag{13}$$

A single server of rate $\mu_1 + \mu_2$ will outperform a 2-server combination of rates $\mu_1$ and $\mu_2$ respectively, and we note from eqn (13) that in a two server system, increasing the rate of the fast server increases the limiting threshold value. This allows us to bound $M_2$ by considering a 2-server combination in which the fast server operates at a rate $\mu_1 + \mu_2$ and the slow server operates at a rate $\mu_3$:

$$M_2 \leq \left\lfloor \frac{\mu_1 + \mu_2}{\mu_3} \right\rfloor \tag{14}$$

Proceeding in a similar fashion yields the following upper bounds on $M_i$:

$$M_i \leq \left\lfloor \mu_{i+1}^{-1} \sum_{j=1}^{i} \mu_j \right\rfloor \quad 1 \leq i \leq n - 1 \tag{15}$$

Lower bounds on $M_i$ can be established in a similar manner, but by pairing $\mu_i$, $i > 1$ with $\mu_1$ in the two-server solution and taking the limit of $\widetilde{M}$ as $\lambda \to \mu_1 + \mu_i$. This follows from observation of the two-server system, noting that $\widetilde{M}$ decreases as $\lambda$ increases, and as $\mu_1$ increases, and that the effect of introducing additional servers could be considered in some sense to increase the effective rate of a single equivalent server. In the two-server combination with $\mu_1$ and $\mu_i$, or, equivalently, $v_1$ and $v_i$, saturation occurs at the intersection of the two lines $v_1 = cv_i$ and $v_1 + v_i = 1$, which occurs at $v_1 = \frac{c}{c+1}$, $v_i = \frac{1}{c+1}$. Substituting these values into the equation for $\widetilde{M}$ yields:

$$M_i \geq \left\lfloor \frac{-1 + \sqrt{4c + 5}}{2} \right\rfloor = \left\lfloor \frac{-1 + \sqrt{4\frac{\mu_1}{\mu_i} + 5}}{2} \right\rfloor \tag{16}$$

Given these bounds on the $M_i$, a finite search process can be used to find the optimal combination of $M_i$'s to minimize $\bar{N}$. More efficient techniques for identifying optimal thresholds for the n-server case are a subject for future investigation.

# CHAPTER 6

## SUMMARY, CONCLUSIONS, AND DIRECTIONS FOR FUTURE RESEARCH

In this dissertation, the dynamic control of multiple exponential servers with dissimilar service rates was considered. The control problem was formulated as a Markovian decision process, then analysis was focussed on the two-server case. A general class of non-preemptive control policies was considered, out of which a specific class of policies, called threshold queueing policies was developed. The steady state performance equations for this class of policies were then formulated from the state transition diagram and an algorithmic solution derived. A major result was the derivation of a closed form solution for the state probabilities for a two-server threshold queueing system in steady state, and the use of this solution to compute $\bar{N}$, the mean number of customers in the system. The equation for $\bar{N}$ was subsequently used as the objective function for the optimization of the control discipline as a function of mean customer arrival rate and mean service rates. It was shown that the form of the optimal threshold queueing discipline is deterministic. This resulted in a substantial simplification of the objective function, which was then used to prove existence and quasi-uniqueness of an optimal threshold size. Finding the optimal threshold size over the operational range of the system parameters was performed using an algorithmic search strategy, and resulted in a partitioned system parameter space, in which the threshold size remains constant within a partition. A linear approximation to the boundaries of the partitions was found which yields a very simple yet good approximation to the optimal threshold size. Utilizing the approximate threshold size was shown numerically to result in less than a 1.5% deterioration in performance (worst case). The overall performance of the threshold queueing discipline was analyzed and compared to a work-conserving discipline, in which a server is never left idle while a customer waits in the queue. It was shown that the degree of improvement provided by threshold queueing is directly related to the ratio of the server rates. The higher this ratio, the more improvement is provided through threshold queueing. No significant improvement is noted until this ratio exceeds two, however.

Several applications and extensions of the threshold queueing discipline were also presented to demonstrate its utility in computer and communications systems. It was shown to be useful for controlling the flow of messages over communication lines of dissimilar capacity, and for managing the use of line printers in a multiprogramming computer system. A time-sharing system example was also developed to demonstrate the applicability of threshold queueing in a system where the slower of two servers is the preferred server. The formulation was extended to treat the finite queue capacity case, and the utility of this particular extension was demonstrated by a spacecraft data management example. Finally, a solution approach to the n-server problem was developed, and upper and lower bounds on the threshold sizes established. As a search strategy is still required to solve the n-server problem, the establishment of these bounds was critical to assuring the finiteness of the search process.

The work presented in this dissertation can be further extended in several directions. The two conjectures made in section 4.4, for example, need to be resolved. The first conjecture addresses the ultimate form of the globally optimal control discipline for the two-server case, and the second restates the conjecture for the n-server case. It is conjectured that the globally optimal discipline is deterministic. There is a lot of evidence suggesting that this is, in fact, the case, both within this dissertation and the operations research literature. The precise proof remains to be constructed, however.

The cost structure employed within this work excluded costs for turning on or off a server, and assumed that customer holding costs and server operation costs were all linear. A more general cost structure would be worth investigating, particularly with regard to its effect on the optimal discipline. Yadin and Naor [YADIM67] pointed out the correlation between these more general cost structures and hysteresis control policies. It could be expected, therefore, that with a more general cost structure, a more complex threshold structure would emerge in which separate thresholds would be required for customer arrivals and departures, respectively.

While the characteristics of the arrival process are, by assumption, well known (Poisson), the nature of the departure process in a threshold queueing system is unknown. While this process may be interesting in its own right, it is required if a threshold queueing server is to be used as a component of a larger model. Similarly, an assumption of exponential service has been made in the analysis presented here. Generalization of both the arrival and service processes should be considered to extend these results to other distributions, and the impact on the departure process considered.

Chapter 4 developed an approximate control law which was based on linear approximations to the boundaries of partitions in the two dimensional parameter space for the two-server problem. This approximation was good, resulting in nearly optimal control. As the problem is extended to the n-server case, the parameter space expands to an n-dimensional hyperspace. Given the increased complexity introduced by multi-dimensionality, suitable approximate control laws may, in a similar fashion, be more effectively sought than exact solutions.

The n-server solution approach outlined in section 5.4 could be used to find a good control law, albeit rather laboriously. Alternatively, and potentially more elegantly, a closed form, general n-server solution could be sought. Due to the exponential growth in complexity of the state transition diagram, however, another solution methodology may be required.

Effectively employing a threshold queueing multi-server within the context of a larger analytic system model implies treating it as one element in a network of queues. Prerequisite to the solution of such a model is the solution to some of the problems outlined above. The analysis of the departure process, for example, was cited. The consideration of non-Poisson arrival distributions may also require treatment.

## APPENDIX A

## DERIVATION OF ALGORITHMIC SOLUTION TO THRESHOLD QUEUEING FOR TWO SERVERS

The steady state equations to be solved are:

a) $p_{0,00} = v_1 p_{0,10} + v_2 p_{0,01}$ (1)

b) $(1 + v_1)p_{0,10} = p_{0,00} + v_1 p_{1,10} + v_2 p_{0,11}$

c) $(1 + v_1)p_{i,10} = p_{i-1,10} + v_1 p_{i+1,10} + v_2 p_{i,11} \quad 1 \leq i \leq M - 1$

d) $(1 + v_2)p_{0,01} = v_1 p_{0,11}$

e) $(1 + v_1 + v_2)p_{0,11} = p_{0,01} + v_1 p_{1,11}$

f) $(1 + v_1 + v_2)p_{i,11} = p_{i-1,11} + v_1 p_{i+1,11} \quad 1 \leq i \leq M - 2$

g) $(1 + v_1 + v_2)p_{M-1,11} = p_{M-2,11} + (1 - \alpha)p_{M-1,10} + [v_1 + (1 - \beta)v_2]\, p_{M,11}$

h) $(1 + v_1)p_{M,10} = \alpha p_{M-1,10} + \beta v_2 p_{M,11}$

i) $(1 + v_1 + v_2)p_{M,11} = p_{M-1,11} + p_{M,10} + (v_1 + v_2)p_{M+1,11}$

j) $p_{>M,11} = \left(\dfrac{1}{v_2 + v_1 - 1}\right) p_{M,11}$

The approach will be to solve each equation in terms of $p_{0,01}$, and then use the normalization condition:

$$p_{0,00} + p_{0,01} + \sum_{i=0}^{M} p_{i,10} + \sum_{i=0}^{\infty} p_{i,11} = 1 \tag{2}$$

to find $p_{0,01}$. To simplify the exposition, we define

$$p_{q,n} = c_{q,n} p_{0,01} \tag{3}$$

and proceed to derive $c_{q,n}$. Eqn (1d) immediately yields:

$$c_{0,11} = \frac{v_2 + 1}{v_1} \tag{4}$$

Eqn (1e) yields:

$$c_{1,11} = \left(\frac{v_2 + v_1 + 1}{v_1}\right) c_{0,11} - \frac{1}{v_1} = \left(\frac{v_2 + 1}{v_1}\right)^2 + \frac{v_2}{v_1} \tag{5}$$

And from eqn (1f):

$$c_{i+1,11} = \left(\frac{v_2 + v_1 + 1}{v_1}\right) c_{i,11} - \frac{1}{v_1} c_{i-1,11} \qquad 1 \leq i \leq M-2 \tag{6}$$

Eqn (1a) can be rewritten:

$$c_{0,10} = \frac{1}{v_1} (c_{0,00} - v_2) \tag{7}$$

Substituting this into eqn (1b) and the resultant into eqn (1c) produces:

$$c_{i,10} = \frac{1}{v_1} \left(c_{i-1,10} - v_2 - v_2 \sum_{j=0}^{i-1} c_{j,11}\right) \qquad 1 \leq i \leq M$$

$$= \left(\frac{1}{v_1}\right)^{i+1} c_{0,00} - \frac{v_2}{v_1} \sum_{j=0}^{i} \left(\frac{1}{v_1}\right)^j - \frac{v_2}{v_1} \sum_{k=0}^{i-1} \sum_{j=0}^{k} \left(\frac{1}{v_1}\right)^j c_{i-k-1,11} \tag{8}$$

Equations (1g) and (1h) can be combined to yield:

$$(1 + v_1)[v_1 + (1 - \beta)v_2]\, c_{M,10} - [\alpha v_1 + (\alpha - \beta)v_2] c_{M-1,10} = \beta v_2 [(v_2 + v_1 + 1)c_{M-1,11} - c_{M-2,11}] \tag{9}$$

Substituting eqn (8) into eqn (9) for $c_{M,10}$ and $c_{M-1,11}$ produces $c_{0,00}$:

$$c_{0,00} = v_1^M \left[1 + \frac{v_2}{v_1}(1 - \beta) + (1 - \alpha)(v_1 + v_2)\right]^{-1} \beta v_2 [(v_2 + v_1 + 1)c_{M-1,11} - c_{M-2,11}]$$

$$+ (1 + v_1)\, [v_1 + (1 - \beta)v_2]\, \frac{v_2}{v_1} \left[\sum_{j=0}^{M} \left(\frac{1}{v_1}\right)^j + \sum_{k=0}^{M-1} \sum_{j=0}^{k} \left(\frac{1}{v_1}\right)^j c_{M-k-1,11}\right]$$

$$- [\alpha v_1 + (\alpha - \beta)v_2]\, \frac{v_2}{v_1} \left[\sum_{j=0}^{M-1} \left(\frac{1}{v_1}\right)^j + \sum_{k=0}^{M-2} \sum_{j=0}^{k} \left(\frac{1}{v_1}\right)^j c_{M-k-2,11}\right] \tag{10}$$

Now from eqn (1h) we get:

$$c_{M,11} = \frac{1}{\beta v_2} [(v_1 + 1)c_{M,10} - \alpha c_{M-1,10}] \tag{11}$$

Since $p_{0,01} = c_{0,01}\, p_{0,01}$, $c_{0,01} = 1$ identically. To solve for $p_{0,01}$, we sum the steady state probabilities:

$$p_{0,00} + p_{0,01} + \sum_{i=0}^{M} p_{i,10} + \sum_{i=0}^{M-1} p_{i,11} + p_{M,11} + p_{>M,11}$$

$$= \left( c_{0,00} + c_{0,01} + \sum_{i=0}^{M} c_{i,10} + \sum_{i=0}^{M-1} c_{i,11} + c_{M,11} + c_{>M,11} \right) p_{0,01} = 1 \qquad (12)$$

## APPENDIX B

## DERIVATION OF CLOSED FORM SOLUTION TO STEADY STATE EQUATIONS FOR TWO SERVERS

In Appendix A the following algorithmic solution was derived for the two server problem:

a) $c_{0,01} = 1$ (1)

b) $c_{0,11} = \dfrac{v_2 + 1}{v_1}$

c) $c_{1,11} = \left(\dfrac{v_2 + 1}{v_1}\right)^2 + \dfrac{v_2}{v_1}$

d) $c_{i,11} = \left(\dfrac{v_2 + v_1 + 1}{v_1}\right) c_{i-1,11} - \dfrac{1}{v_1} c_{i-2,11} \qquad 2 \leq i \leq M - 1$

e) $c_{0,00} = v_1^M \left[1 + \dfrac{v_2}{v_1}(1 - \beta) + (1 - \alpha)(v_2 + v_1)\right]^{-1} \left\{\beta v_2 [(v_2 + v_1 + 1)c_{M-1,11} - c_{M-2,11}]\right.$

$$+ (1 + v_1)[v_1 + (1 - \beta)v_2] \frac{v_2}{v_1}\left[\sum_{j=0}^{M}\left(\frac{1}{v_1}\right)^j + \sum_{k=0}^{M-1}\sum_{j=0}^{k}\left(\frac{1}{v_1}\right)^j c_{M-k-1,11}\right]$$

$$\left. - [\alpha v_1 + (\alpha - \beta)v_2] \frac{v_2}{v_1}\left[\sum_{j=0}^{M-1}\left(\frac{1}{v_1}\right)^j + \sum_{k=0}^{M-2}\sum_{j=0}^{k}\left(\frac{1}{v_1}\right)^j c_{M-k-2,11}\right]\right\}$$

f) $c_{0,10} = \dfrac{1}{v_1}(c_{0,00} - v_2)$

g) $c_{i,10} = \dfrac{1}{v_1}\left(c_{i-1,10} - v_2 - v_2 \sum_{j=0}^{i-1} c_{j,11}\right) \qquad 1 \leq i \leq M$

h) $c_{M,11} = \dfrac{1}{\beta v_2}[(v_1 + 1)c_{M,10} - \alpha c_{M-1,10}]$

i) $c_{M+i,11} = \left(\dfrac{1}{v_2 + v_1}\right)^i c_{M,11}$

j) $c_{>M,11} = \left(\dfrac{1}{v_2 + v_1 - 1}\right) c_{M,11}$

k) $$p_{0,01} = \left(c_{0,00} + c_{0,01} + \sum_{i=0}^{M} c_{i,10} + \sum_{i=0}^{M-1} c_{i,11} + c_{M,11} + c_{>M,11}\right)^{-1}$$

l) $p_{q,n} = c_{q,n} p_{0,01}$ for all states (q, n)

In this appendix, the following solution is derived from eqns (1a – l):

a) $$c_{i,11} = f_{i+1} - f_i \qquad 0 \le i \le M-1 \tag{2}$$

b) $$c_{0,00} = \frac{v_1^M v_2 \{-(v_2+v_1)g_{M-1}\alpha + [v_1(f_M - f_{M-1}) - v_2 g_M]\beta + (v_1+1)(v_2+v_1)g_M\}}{-v_1(v_2+v_1)\alpha - v_2\beta + (v_1+1)(v_2+v_1)}$$

c) $$c_{i,10} = \frac{\begin{array}{r} v_2(v_2+v_1)(g_i - v_1^{M-1-i} g_{M-1})\alpha + \left[v_1^{M-1-i} v_2(v_1 f_M - v_1 f_{M-1} - v_2 g_M) + \frac{v_2^2}{v_1} g_i\right]\beta \\ + v_2(v_1+1)(v_2+v_1)\left(v_1^{M-1-i} g_M - \frac{1}{v_1} g_i\right) \end{array}}{-v_1(v_2+v_1)\alpha - v_2\beta + (v_1+1)(v_2+v_1)} \qquad 0 \le i \le M$$

d) $$c_{M,11} = \frac{-[(v_1 - v_2)f_M - v_1 f_{M-1}]\alpha + (v_1+1)(f_M - f_{M-1})}{-v_1(v_2+v_1)\alpha - v_2\beta + (v_1+1)(v_2+v_1)}$$

e) $$c_{M+i,11} = \left(\frac{1}{v_2+v_1}\right)^i c_{M,11}$$

f) $$\sum_{i=0}^{\infty} c_{M+i,11} = \left(\frac{v_2+v_1}{v_2+v_1-1}\right) c_{M,11}$$

g) $$p_{0,01}^{-1} = \frac{\begin{array}{r} -(v_2+v_1)\left[v_2(v_2+v_1-1)\sum_{i=0}^{M-1}\sum_{j=0}^{i} v_1^{j+1} f_i + (v_1^2 - v_2^2)f_M - v_1 f_{M-1}\right]\alpha \\ -v_2(v_2+v_1-1)\left[v_2\sum_{i=0}^{M}\sum_{j=0}^{i} v_1^j f_i + f_{M-1}\sum_{i=0}^{M+1} v_1^i - f_M\sum_{i=1}^{M+1} v_1^i\right]\beta \\ + (v_1+1)(v_2+v_1)\left[(v_2+v_1-1)\left(f_M + v_2\sum_{i=0}^{M}\sum_{j=0}^{i} v_1^j f_i\right) + f_M - f_{M-1}\right] \end{array}}{(v_2+v_1-1)[-v_1(v_2+v_1)\alpha - v_2\beta + (v_1+1)(v_2+v_1)]}$$

where $f_i$ and $g_k$ denote $f_i\left(\frac{v_2+v_1+1}{v_1}, \frac{1}{v_1}\right)$ and $g_k\left(\frac{v_2+v_1+1}{v_1}, \frac{1}{v_1}\right)$, respectively, and:

$$f_i(x, y) = \sum_{j=0}^{\lfloor \frac{i}{2} \rfloor} (-1)^j \binom{i-j}{j} x^{i-2j} y^j \tag{3}$$

$$g_k(x, y) = \sum_{i=0}^{k} y^{k-i} f_i(x, y) \tag{4}$$

The derivation begins with eqn (1d), for which the following result is fundamental:

Theorem B.1 If $\theta_i = x\theta_{i-1} - y\theta_{i-2}$ for $i \geq 2$, then $\theta_i = \theta_1 f_{i-1}(x, y) - y\theta_0 f_{i-2}(x, y)$, where $f_i(x, y)$ is given by eqn (3).

Proof The proof proceeds by induction on i. To prove the initial result, it is clear that $\theta_2 = \theta_1 f_1(x, y) - y\theta_0 f_0(x, y) = x\theta_1 - y\theta_0$, as required. Now assume the relation is true for $\theta_i$, then for $\theta_{i+1}$ we have:

$$\theta_{i+1} = x\theta_i - y\theta_{i-1} = [xf_{i-1}(x, y) - yf_{i-2}(x, y)]\theta_1 - y[xf_{i-2}(x, y) - yf_{i-3}(x, y)]\theta_0 \tag{5}$$

and we wish to show that this is equivalent to:

$$\theta_1 f_i(x, y) - y\theta_0 f_{i-1}(x, y).$$

The equivalence certainly holds if it can be shown that:

$$f_i(x, y) = xf_{i-1}(x, y) - yf_{i-2}(x, y) \tag{6}$$

Substituting eqn (3) for $f_i(x, y)$ and rearranging, eqn (6) becomes:

$$\sum_{j=0}^{\lfloor \frac{i-1}{2} \rfloor} (-1)^j \binom{i-1-j}{j} x^{i-2j} y^j - \sum_{j=0}^{\lfloor \frac{i}{2} \rfloor} (-1)^j \binom{i-j}{j} x^{i-2j} y^j = \sum_{j=0}^{\lfloor \frac{i-2}{2} \rfloor} (-1)^j \binom{i-2-j}{j} x^{i-2-2j} y^{j+1} \tag{7}$$

The cases in which i is even or odd will be considered separately. Consider first i being odd, for which $\left\lfloor \frac{i-1}{2} \right\rfloor = \left\lfloor \frac{i}{2} \right\rfloor$. Then eqn (7) can be written:

$$-\sum_{j=1}^{\lfloor \frac{i}{2} \rfloor} (-1)^j \binom{i-1-j}{j-1} x^{i-2j} y^j = \sum_{j=0}^{\lfloor \frac{i-2}{2} \rfloor} (-1)^j \binom{i-2-j}{j} x^{i-2-2j} y^{j+1} \tag{8}$$

where the relation $\binom{n}{m} + \binom{n}{m+1} = \binom{n+1}{m+1}$ has been used to combine terms. Substituting $j - 1$ for j on the left side of eqn (8) and adjusting the limits of summation accordingly

proves the relation. Now for the case where i is even, the left side of eqn (7) can be written:

$$\sum_{j=0}^{\left\lfloor\frac{i-1}{2}\right\rfloor}(-1)^j\binom{i-1-j}{j}x^{i-2j}y^j-\sum_{j=0}^{\left\lfloor\frac{i-1}{2}\right\rfloor}(-1)^j\binom{i-j}{j}x^{i-2j}y^j-(-1)^{\frac{i}{2}}y^{\frac{i}{2}}$$

$$=\sum_{j=1}^{\left\lfloor\frac{i-1}{2}\right\rfloor}(-1)^{j-1}\binom{i-1-j}{j-1}x^{i-2j}y^j-(-1)^{\frac{i}{2}}y^{\frac{i}{2}}$$

$$=\sum_{j=0}^{\left\lfloor\frac{i-3}{2}\right\rfloor}(-1)^j\binom{i-2-j}{j}x^{i-2(j+1)}y^{j+1}+(-1)^{\left\lfloor\frac{i-2}{2}\right\rfloor}\binom{i-2-\left\lfloor\frac{i-2}{2}\right\rfloor}{\left\lfloor\frac{i-2}{2}\right\rfloor}y^{\left\lfloor\frac{i-2}{2}\right\rfloor+1}$$

$$=\sum_{j=0}^{\left\lfloor\frac{i-2}{2}\right\rfloor}(-1)^j\binom{i-2-j}{j}x^{i-2-2j}y^{j+1}$$

This completes the proof of theorem B.1.

Applying this result to eqn (1d) results in the following:

$$c_{i,11}=\left[\left(\frac{v_2+1}{v_1}\right)^2+\frac{v_2}{v_1}\right]f_{i-1}-\left(\frac{v_2+1}{v_1^2}\right)f_{i-2}\qquad 2\le i\le M-1 \tag{9}$$

Now, from eqn (6), with $x=\dfrac{v_2+v_1+1}{v_1}$ and $y=\dfrac{1}{v_1}$, we have:

$$f_i=\left(\frac{v_2+v_1+1}{v_1}\right)f_{i-1}-\left(\frac{1}{v_1}\right)f_{i-2} \tag{10}$$

Substituting for $f_{i-2}$ in eqn (9) yields:

$$c_{i,11}=\left(\frac{v_2+1}{v_1}\right)f_i-\frac{1}{v_1}f_{i-1}\qquad 2\le i\le M-1 \tag{11}$$

Carrying out this same substitution, but now for $f_{i-1}$ yields:

$$c_{i,11}=f_{i+1}-f_i\qquad 2\le i\le M-1 \tag{12}$$

Calculating the first three terms of $f_i$:

$$\text{a) } f_0\left(\frac{v_2+v_1+1}{v_1}, \frac{1}{v_1}\right) = 1 \tag{13}$$

$$\text{b) } f_1\left(\frac{v_2+v_1+1}{v_1}, \frac{1}{v_1}\right) = \frac{v_2+v_1+1}{v_1}$$

$$\text{c) } f_2\left(\frac{v_2+v_1+1}{v_1}, \frac{1}{v_1}\right) = \left(\frac{v_2+v_1+1}{v_1}\right)^2 - \frac{1}{v_1}$$

Then we have:

$$\text{a) } f_1 - f_0 = \frac{v_2+1}{v_1} = c_{0,11} \tag{14}$$

$$\text{b) } f_2 - f_1 = \left(\frac{v_2+1}{v_1}\right)^2 + \frac{v_2}{v_1} = c_{1,11}$$

And it is clear that the range of eqn (12) can be extended:

$$c_{i,11} = f_{i+1} - f_i \quad 0 \leq i \leq M-1 \tag{15}$$

which corresponds to eqn (2a).

Now turning to eqn (1e) for $c_{0,00}$, the first step is to simplify the subexpression:

$$\sum_{j=0}^{M}\left(\frac{1}{v_1}\right)^j + \sum_{k=0}^{M-1}\sum_{j=0}^{k}\left(\frac{1}{v_1}\right)^j c_{M-k-1,11}$$

$$= \left(1 + \sum_{j=0}^{M-1} c_{j,11}\right) + \frac{1}{v_1}\left(1 + \sum_{j=0}^{M-2} c_{j,11}\right) + \ldots + \frac{1}{v_1^{M-1}}(1 + c_{0,11}) + \frac{1}{v_1^M}$$

$$= f_M + \frac{1}{v_1} f_{M-1} + \ldots + \frac{1}{v_1^{M-1}} f_1 + \frac{1}{v_1^M} f_0$$

$$= \sum_{j=0}^{M}\left(\frac{1}{v_1}\right)^{M-j} f_j\left(\frac{v_2+v_1+1}{v_1}, \frac{1}{v_1}\right)$$

$$= g_M\left(\frac{v_2+v_1+1}{v_1}, \frac{1}{v_1}\right) \tag{16}$$

Then eqn (1e) can be written:

$$c_{0,00} = \frac{v_1^M\{\beta v_1 v_2[(v_2 + v_1 + 1)c_{M-1,11} - c_{M-2,11}] - v_2[(\alpha-\beta)v_2 + \alpha v_1]g_{M-1} + v_2(v_1+1)[(1-\beta)v_2 + v_1]g_M\}}{-v_1(v_2+v_1)\alpha - v_2\beta + (v_1+1)(v_2+v_1)}$$

$$= \frac{\begin{array}{l} v_1^M v_2\{-(v_2+v_1)g_{M-1}\alpha + [-v_2(v_1+1)g_M + v_2 g_{M-1} + v_1(v_2+v_1+1)c_{M-1,11} - v_1 c_{M-2,11}]\beta \\ \qquad + (v_1+1)(v_2+v_1)g_M\} \end{array}}{-v_1(v_2+v_1)\alpha - v_2\beta + (v_1+1)(v_2+v_1)} \qquad (17)$$

Noting that from the definition of $g_k$ the following holds:

$$v_1 g_k - g_{k-1} = v_1 f_k \qquad (18)$$

and that from eqns (15) and (10):

$$\begin{aligned} (v_2+v_1)c_{i,11} - c_{i-1,11} &= (v_2+v_1)(f_{i+1} - f_i) - (f_i - f_{i-1}) \qquad 1 \le i \le M-1 \\ &= (v_2+v_1)f_{i+1} - (v_2+v_1+1)f_i + f_{i-1} \\ &= v_2 f_{i+1} \end{aligned} \qquad (19)$$

then eqn (17) can be written:

$$c_{0,00} = \frac{v_1^M v_2\{-(v_2+v_1)g_{M-1}\alpha + (v_1 c_{M-1,11} - v_2 g_M)\beta + (v_1+1)(v_2+v_1)g_M\}}{-v_1(v_2+v_1)\alpha - v_2\beta + (v_1+1)(v_2+v_1)}$$

$$= \frac{v_1^M v_2\{-(v_2+v_1)g_{M-1}\alpha + [v_1(f_M - f_{M-1}) - v_2 g_M]\beta + (v_1+1)(v_2+v_1)g_M\}}{-v_1(v_2+v_1)\alpha - v_2\beta + (v_1+1)(v_2+v_1)} \qquad (20)$$

which corresponds to eqn (2b).

Now turning to eqns (1g) and substituting for $c_{j,11}$:

$$c_{i,10} = \frac{1}{v_1}(c_{i-1,10} - v_2 f_i) \qquad 1 \le i \le M \qquad (21)$$

which can be written:

$$c_{i,10} = \left(\frac{1}{v_1}\right)^i c_{0,10} - \frac{v_2}{v_1} \sum_{j=1}^{i} \left(\frac{1}{v_1}\right)^{i-j} f_j \qquad 1 \le i \le M \tag{22}$$

Substituting for $c_{0,10}$ from eqn (1f):

$$c_{i,10} = \left(\frac{1}{v_1}\right)^{i+1} c_{0,00} - \frac{v_2}{v_1} \sum_{j=0}^{i} \left(\frac{1}{v_1}\right)^{i-j} f_j \qquad 1 \le i \le M$$

$$= \left(\frac{1}{v_1}\right)^{i+1} c_{0,00} - \frac{v_2}{v_1}\, g_i \left(\frac{v_2 + v_1 + 1}{v_1}, \frac{1}{v_1}\right) \tag{23}$$

The range of eqn (23) can be extended to include $c_{0,10}$, as can readily be verified by inspection of eqn (1f):

$$c_{i,10} = \left(\frac{1}{v_1}\right)^{i+1} c_{0,00} - \frac{v_2}{v_1}\, g_i \left(\frac{v_2 + v_1 + 1}{v_1}, \frac{1}{v_1}\right) \qquad 0 \le i \le M \tag{24}$$

Substituting eqn (20) for $c_{0,00}$ into eqn (24) and simplifying yields:

$$c_{i,10} = \frac{\begin{array}{l} v_2(v_2 + v_1)(g_i - v_1^{M-1-i} g_{M-1})\alpha \\ \quad + \left[v_1^{M-i-1} v_2 (v_1 f_M - v_1 f_{M-1,11} - v_2 g_M) + \dfrac{v_2^2}{v_1} g_i\right]\beta \\ \quad + v_2(v_1 + 1)(v_2 + v_1)\left(v_1^{M-1-i} g_M - \dfrac{1}{v_1} g_i\right) \end{array}}{-v_1(v_2 + v_1)\alpha - v_2\beta + (v_1 + 1)(v_2 + v_1)} \qquad 0 \le i \le M \tag{25}$$

which corresponds to eqn (2c).

Now from eqn (1h) for $c_{M,11}$ and eqn (25) we get eqn (2d):

$$c_{M,11} = \frac{\begin{array}{l} -\left\{v_2[-(v_1 + 1)g_M + g_{M-1}] + \dfrac{v_2}{v_1} g_{M-1} + v_1[(v_2 + v_1 + 1)c_{M-1,11} - c_{M-2,11}]\right\}\alpha \\ \quad + (v_1 + 1)\left[\dfrac{v_2}{v_1}(-v_1 g_M + g_{M-1}) + (v_2 + v_1 + 1)c_{M-1,11} - c_{M-2,11}\right] \end{array}}{-v_1(v_2 + v_1)\alpha - v_2\beta + (v_1 + 1)(v_2 + v_1)}$$

$$- \left[ v_2(-v_1 f_M - g_M) + \frac{v_2}{v_1} g_{M-1} + v_1(v_2 f_M + c_{M-1,11}) \right] \alpha$$

$$= \frac{+ (v_1 + 1)(-v_2 f_M + v_2 f_M + c_{M-1,11})}{-v_1(v_2 + v_1)\alpha - v_2\beta + (v_1 + 1)(v_2 + v_1)}$$

$$= \frac{-[(v_1 - v_2)f_M - v_1 f_{M-1}]\alpha + (v_1 + 1)(f_M - f_{M-1})}{-v_1(v_2 + v_1)\alpha - v_2\beta + (v_1 + 1)(v_2 + v_1)} \tag{26}$$

For this system to be ergodic, $v_2 + v_1 > 1$, and so eqn (2e) yields eqn (2f) directly:

$$\sum_{i=0}^{\infty} c_{M+i,11} = \sum_{i=0}^{\infty} \left( \frac{1}{v_2 + v_1} \right)^i c_{M,11}$$

$$= \frac{1}{1 - \left( \frac{1}{v_2 + v_1} \right)} c_{M,11}$$

$$= \frac{v_2 + v_1}{v_2 + v_1 - 1} c_{M,11} \tag{27}$$

The remaining equation to develop is (2g), for $p_{0,01}$. The first step in this process is to evaluate $\sum_{i=0}^{M} c_{i,10}$.

$$\sum_{i=0}^{M} c_{i,10} = \sum_{i=0}^{M} \frac{\begin{array}{l} v_2(v_2 + v_1)(g_i - v_1^{M-1-i} g_{M-1})\alpha + \left[ v_1^{M-1-i} v_2(v_1 f_M - v_1 f_{M-1} - v_2 g_M) + \frac{v_2^2}{v_1} g_i \right] \beta \\ + v_2(v_1 + 1)(v_2 + v_1) \left( v_1^{M-1-i} g_M - \frac{1}{v_1} g_i \right) \end{array}}{-v_1(v_2 + v_1)\alpha - v_2\beta + (v_1 + 1)(v_2 + v_1)} \tag{28}$$

Consider the coefficient of $\alpha$ in the numerator.

$$\sum_{i=0}^{M} v_2(v_2 + v_1)(g_i - v_1^{M-1-i} g_{M-1}) = v_2(v_2 + v_1) \left[ g_0 + g_1 + \ldots + g_M - \left( v_1^{M-1} + v_1^{M-2} + \ldots + 1 + \frac{1}{v_1} \right) g_{M-1} \right]$$

$$= v_2(v_2 + v_1) \left[ f_0 + \left( \frac{1}{v_1} f_0 + f_1 \right) + \ldots + \left( \frac{1}{v_1^M} f_0 + \ldots + f_M \right) \right.$$

$$- \left( v_1^{M-1} + v_1^{M-2} + \ldots + 1 + \frac{1}{v_1} \right) \left( \frac{1}{v_1^{M-1}} f_0 + \frac{1}{v_1^{M-2}} f_1 + \ldots + f_{M-1} \right) \Bigg]$$

$$= v_2(v_2 + v_1)[-v_1 f_1 - (v_1^2 + v_1) f_2 - \ldots - (v_1^{M-1} + v_1^{M-2} + \ldots + v_1) f_{M-1} + f_M]$$

$$= v_2(v_2 + v_1) \left[ f_M - \sum_{i=1}^{M-1} \sum_{j=1}^{i} v_1^j f_i \right] \tag{29}$$

A similar simplification process for the third (non – $\alpha$, $\beta$) term in the numerator yields:

$$\sum_{i=0}^{M} v_2(v_1 + 1)(v_2 + v_1) \left( v_1^{M-1-i} g_M - \frac{1}{v_1} g_i \right) = \frac{v_2}{v_1} (v_1 + 1)(v_2 + v_1) \sum_{i=1}^{M} \sum_{j=1}^{i} v_1^j f_i \tag{30}$$

so that eqn (28) becomes:

$$\sum_{i=0}^{M} c_{i,10} = \frac{\begin{array}{r} v_2(v_2 + v_1) \left[ f_M - \sum_{i=1}^{M-1} \sum_{j=1}^{i} v_1^j f_i \right] \alpha + \frac{v_2}{v_1} \sum_{i=0}^{M} [v_1^i (v_1 f_M - v_1 f_{M-1} - v_2 g_M) + v_2 g_i] \beta \\ + \frac{v_2}{v_1} (v_1 + 1)(v_2 + v_1) \sum_{i=1}^{M} \sum_{j=1}^{i} v_1^j f_i \end{array}}{-v_1(v_2 + v_1)\alpha - v_2\beta + (v_1 + 1)(v_2 + v_1)} \tag{31}$$

Now for $p_{0,01}$, we have:

$$p_{0,01}^{-1} = c_{0,00} + c_{0,01} + \sum_{i=0}^{M} c_{i,10} + \sum_{i=0}^{M-1} c_{i,11} + \sum_{i=0}^{\infty} c_{M+i,11}$$

$$= \frac{v_1^M v_2 \{-(v_2 + v_1) g_{M-1} \alpha + [v_1(f_M - f_{M-1}) - v_2 g_M] \beta + (v_1 + 1)(v_2 + v_1) g_M\}}{-v_1(v_2 + v_1)\alpha - v_2\beta + (v_1 + 1)(v_2 + v_1)}$$

$$+ 1 + \frac{v_2(v_2 + v_1) \left[ f_M - \sum_{i=1}^{M-1} \sum_{j=1}^{i} v_1^j f_i \right] \alpha + \frac{v_2}{v_1} \sum_{i=0}^{M} [v_1^i (v_1 f_M - v_1 f_{M-1} - v_2 g_M) + v_2 g_i] \beta}{-v_1(v_2 + v_1)\alpha - v_2\beta + (v_1 + 1)(v_2 + v_1)}$$

$$+ \frac{\frac{v_2}{v_1} (v_1 + 1)(v_2 + v_1) \sum_{i=1}^{M} \sum_{j=1}^{i} v_1^j f_i}{-v_1(v_2 + v_1)\alpha - v_2\beta + (v_1 + 1)(v_2 + v_1)} - 1 + f_M$$

$$- \left( \frac{v_2 + v_1}{v_2 + v_1 - 1} \right) \frac{[(v_1 - v_2) f_M - v_1 f_{M-1}] \alpha - (v_1 + 1)(f_M - f_{M-1})}{-v_1(v_2 + v_1)\alpha - v_2\beta + (v_1 + 1)(v_2 + v_1)}$$

$$\left\{(v_2+v_1-1)(v_2+v_1)\left[-v_1^M v_2 g_{M-1} - v_2\sum_{i=1}^{M-1}\sum_{j=1}^{i} v_1^j f_i - v_1 f_M\right] - (v_2+v_1)[v_1 c_{M-1,11} - (v_2+v_1)v_2 f_M]\right\}\alpha$$

$$+(v_2+v_1-1)\left[v_2 c_{M-1,11}\sum_{i=0}^{M+1} v_1^i - v_1^M v_2^2 g_M - \frac{v_2^2}{v_1}\sum_{i=1}^{M}\sum_{j=1}^{i} v_1^j f_i - v_2 f_M\right]\beta$$

$$+(v_1+1)(v_2+v_1)\left\{(v_2+v_1-1)\left[v_1^M v_2 g_M + \frac{v_2}{v_1}\sum_{i=1}^{M}\sum_{j=1}^{i} v_1^j f_i + f_M\right] + c_{M-1,11}\right\}$$

$$= \frac{}{(v_2+v_1-1)[-v_1(v_2+v_1)\alpha - v_2\beta + (v_1+1)(v_2+v_1)]} \quad (32)$$

Utilizing the following identity, eqn (32) can be further simplified.

$$v_1^M g_M + \sum_{i=1}^{M}\sum_{j=1}^{i} v_1^{j-1} f_i = \sum_{j=0}^{M} v_1^j f_j + \sum_{i=1}^{M}\sum_{j=1}^{i} v_1^{j-1} f_i$$

$$= f_0 + (1+v_1)f_1 + (1+v_1+v_1^2)f_2 + \ldots + (1+v_1+\ldots+v_1^M)f_M$$

$$= \sum_{i=0}^{M}\sum_{j=0}^{i} v_1^j f_i \quad (33)$$

Performing this substitution yields:

$$-(v_2+v_1)\left\{(v_2+v_1-1)v_1\left[f_M + v_2\sum_{i=0}^{M-1}\sum_{j=0}^{i} v_1^j f_i\right] + v_1 c_{M-1,11} - (v_2+v_1)v_2 f_M\right\}\alpha$$

$$-(v_2+v_1-1)v_2\left[f_M + v_2\sum_{i=0}^{M}\sum_{j=0}^{i} v_1^j f_i - c_{M-1,11}\sum_{i=0}^{M+1} v_1^i\right]\beta$$

$$+(v_1+1)(v_2+v_1)\left\{(v_2+v_1-1)\left[f_M + v_2\sum_{i=0}^{M}\sum_{j=0}^{i} v_1^j f_i\right] + c_{M-1,11}\right\}$$

$$p_{0,01}^{-1} = \frac{}{(v_2+v_1-1)[-v_1(v_2+v_1)\alpha - v_2\beta + (v_1+1)(v_2+v_1)]}$$

$$
\begin{aligned}
&-(v_2+v_1)\left[v_2(v_2+v_1-1)\sum_{i=0}^{M-1}\sum_{j=0}^{i}v_1^{j+1}f_i+(v_1^2-v_2^2)f_M-v_1f_{M-1}\right]\alpha\\
&\quad-v_2(v_2+v_1-1)\left[v_2\sum_{i=0}^{M}\sum_{j=0}^{i}v_1^{j}f_i+f_{M-1}\sum_{i=0}^{M+1}v_1^{i}-f_M\sum_{i=1}^{M+1}v_1^{i}\right]\beta\\
&\quad+(v_1+1)(v_2+v_1)\left[(v_2+v_1-1)\left(f_M+v_2\sum_{i=0}^{M}\sum_{j=0}^{i}v_1^{j}f_i\right)+f_M-f_{M-1}\right]\\
&=\frac{}{(v_2+v_1-1)[-v_1(v_2+v_1)\alpha-v_2\beta+(v_1+1)(v_2+v_1)]}
\end{aligned}
\tag{34}
$$

which is the desired result.

## APPENDIX C

## DERIVATION OF MEAN NUMBER OF CUSTOMERS IN THE SYSTEM ($\bar{N}$)

The mean number of customers in the system ($\bar{N}$) is given by the sum over the state space of the product of the number of customers in the system when the system is in a given state and the steady state probability of the system being in that state:

$$\bar{N} = p_{0,01} + \sum_{i=0}^{M} (i+1)p_{i,10} + \sum_{i=0}^{\infty} (i+2)p_{i,11}$$

$$= p_{0,01}\left[1 + \sum_{i=0}^{M} (i+1)c_{i,10} + \sum_{i=0}^{M-1} (i+2)c_{i,11} + \sum_{i=0}^{\infty} (M+2+i)c_{M+i,11}\right]$$

$$= p_{0,01}\left[1 + \sum_{i=0}^{M} (i+1)c_{i,10} + \sum_{i=0}^{M-1} (i+2)c_{i,11} + \sum_{i=0}^{\infty} (M+2+i)\left(\frac{1}{v_2+v_1}\right)^i c_{M,11}\right]$$

$$= p_{0,01}\left[1 + \sum_{i=0}^{M} (i+1)c_{i,10} + \sum_{i=0}^{M-1} (i+2)c_{i,11} + \left(\frac{v_2+v_1}{v_2+v_1-1}\right)\left(M+1+\frac{v_2+v_1}{v_2+v_1-1}\right)c_{M,11}\right] \quad (1)$$

The term $\sum_{i=0}^{M} (i+1)c_{i,10}$ becomes:

$$\sum_{i=0}^{M} (i+1)c_{i,10} = \frac{\begin{array}{l} \sum_{i=0}^{M} (i+1)v_2(v_2+v_1)(g_i - v_1^{M-1-i}g_{M-1})\alpha \\ + \sum_{i=0}^{M} (i+1)\left[v_1^{M-1-i}v_2(v_1 f_M - v_1 f_{M-1} - v_2 g_M) + \frac{v_2^2}{v_1} g_i\right]\beta \\ + v_2(v_1+1)(v_2+v_1)\sum_{i=0}^{M} (i+1)\left(v_1^{M-1-i}g_M - \frac{1}{v_1} g_i\right) \end{array}}{-v_1(v_2+v_1)\alpha - v_2\beta + (v_1+1)(v_2+v_1)}$$

$$= \frac{\begin{array}{l} v_2(v_2+v_1)\sum_{i=0}^{M} (i+1)\left[\sum_{j=0}^{i} v_1^{j-i}f_j - \sum_{j=0}^{M-1} v_1^{j-i}f_j\right]\alpha \\ + \sum_{i=0}^{M} (i+1)v_2\left[v_1^{M-i}(f_M - f_{M-1}) - v_2\sum_{j=0}^{M} v_1^{j-1-i}f_j + v_2\sum_{j=0}^{i} v_1^{j-1-i}f_j\right]\beta \\ + v_2(v_1+1)(v_2+v_1)\sum_{i=0}^{M} (i+1)\left[\sum_{j=0}^{M} v_1^{j-1-i}f_j - \sum_{j=0}^{i} v_1^{j-1-i}f_j\right] \end{array}}{-v_1(v_2+v_1)\alpha - v_2\beta + (v_1+1)(v_2+v_1)}$$

$$v_2(v_2+v_1)\left[-\sum_{i=1}^{M-1} i \sum_{j=i}^{M-1} v_1^{j-i+1} f_j + (M+1)f_M\right]\alpha$$

$$+\left[\sum_{i=1}^{M+1} i v_2 v_1^{M-i+1}(f_M - f_{M-1}) - \sum_{i=1}^{M} i v_2^2 \sum_{j=i}^{M} v_1^{j-i} f_j\right]\beta$$

$$+ v_2(v_1+1)(v_2+v_1)\sum_{i=1}^{M} i \sum_{j=i}^{M} v_1^{j-i} f_j$$

$$= \frac{}{-v_1(v_2+v_1)\alpha - v_2\beta + (v_1+1)(v_2+v_1)} \tag{2}$$

The term $\sum_{i=0}^{M-1} (i+2)c_{i,11}$ becomes:

$$\sum_{i=0}^{M-1} (i+2)c_{i,11} = \sum_{i=0}^{M-1} (i+2)(f_{i+1} - f_i)$$

$$= 2(f_1 - f_0) + 3(f_2 - f_1) + \ldots + (M+1)(f_M - f_{M-1})$$

$$= -1 - \sum_{i=0}^{M-1} f_i + (M+1)f_M \tag{3}$$

Substituting into eqn (1), we can write:

$$\left[-\sum_{i=0}^{M-1} f_i + (M+1)f_M\right]\left[-v_1(v_2+v_1)\alpha - v_2\beta + (v_1+1)(v_2+v_1)\right]$$

$$+ v_2(v_2+v_1)\left[-\sum_{i=1}^{M-1} i \sum_{j=i}^{M-1} v_1^{j-i+1} f_j + (M+1)f_M\right]\alpha$$

$$+\left[\sum_{i=1}^{M+1} i v_2 v_1^{M-i+1}(f_M - f_{M-1}) - \sum_{i=1}^{M} i v_2^2 \sum_{j=i}^{M} v_1^{j-i} f_j\right]\beta$$

$$+ v_2(v_1+1)(v_2+v_1)\sum_{i=1}^{M} i \sum_{j=i}^{M} v_1^{j-i} f_j$$

$$\frac{\bar{N}}{p_{0,01}} = \frac{+\left(\frac{v_2+v_1}{v_2+v_1-1}\right)\left(M+1+\frac{v_2+v_1}{v_2+v_1-1}\right)\left\{[(v_1-v_2)f_M - v_1 f_{M-1}]\alpha - v_1(v_1+1)(f_M - f_{M-1})\right\}}{-v_1(v_2+v_1)\alpha - v_2\beta + (v_1+1)(v_2+v_1)} \tag{4}$$

Substitution of eqn (B–34) for $p_{0,01}$ and straightforward simplification yields eqns (4–3) and (4–4).

## APPENDIX D

## DERIVATION OF RELATIONSHIPS AMONG $H_i^M(v_1, v_2)$

Three equalities are utilized in section 4.1, stated in eqn (4.1–6), and restated here, where $H_i \equiv H_i^M(v_1, v_2)$:

a) $H_1H_6 - H_3H_4 = (v_1 + 1)(v_2 + v_1 - 1)(v_2 + v_1)^2 f_M F_M(v_1, v_2)$ (1)

b) $H_2H_6 - H_3H_5 = (v_1 + 1)(v_2 + v_1 - 1)(v_2 + v_1)(f_M - f_{M-1})F_M(v_1, v_2)$

c) $H_1H_5 - H_2H_4 = (v_2 + v_1 - 1)(v_2 + v_1)[(v_1 - v_2)f_M - v_1 f_{M-1}]F_M(v_1, v_2)$

where

$$F_M(v_1, v_2) = v_2\left[(M + 2)(v_2 + v_1 - 1)(v_2 + v_1) + 1 + (v_2 + v_1 - 1)^2\left(v_2\sum_{i=1}^{M} j v_1^{M-j} - 1\right)\right]\sum_{i=0}^{M-1}\sum_{j=0}^{i} v_1^{j+1} f_i$$

$$- v_1(v_2 + v_1 - 1)\left[1 + (v_2 + v_1 - 1)\left(v_2\sum_{j=0}^{M} v_1^j + 1\right)\right]\left(v_2\sum_{i=1}^{M-1}\sum_{j=i}^{M-1} i v_1^{j-i} f_j - \sum_{i=0}^{M-2} f_i\right)$$

$$- v_2(v_1^2 - v_2^2)[(M + 1)(v_2 + v_1 - 1) + 1]\sum_{i=0}^{M}\sum_{j=0}^{i} v_1^j f_i$$

$$+ (v_1^2 - v_2^2)(v_2 + v_1 - 1)\left(v_2\sum_{i=1}^{M}\sum_{j=i}^{M} i v_1^{j-i} f_j - \sum_{i=0}^{M-1} f_i\right)$$

$$+ \left\{v_1 v_2[(v_2 + v_1)^2 + M(v_2 + v_1 - 1)]\sum_{j=0}^{M} v_1^j - v_1 v_2(v_2 + v_1 - 1)\sum_{j=1}^{M} j v_1^{M-j}\right.$$

$$\left. + v_1(v_2 + v_1)^2 - v_1^2 + v_2^2\right\} f_{M-1} \qquad (2)$$

and

a) $$H_1^M(v_1, v_2) = -(v_2 + v_1)\left\{(v_2 + v_1 - 1)^2\left[v_2\sum_{i=1}^{M-1}\sum_{j=i}^{M-1} i v_1^{j-i+1} f_j - v_1\sum_{i=0}^{M-1} f_i\right]\right.$$
$$\left. + (v_1^2 - v_2^2)[(M + 1)(v_2 + v_1 - 1) + 1] f_M - v_1[(M + 2)(v_2 + v_1 - 1) + 1] f_{M-1}\right\} \qquad (3)$$

b) $$H_2^M(v_1, v_2) = -v_2(v_2 + v_1 - 1)^2\left\{v_2\sum_{i=1}^{M}\sum_{j=i}^{M} i v_1^{j-i} f_j + (f_{M-1} - f_M)\sum_{i=1}^{M} i v_1^{M-i+1} - \sum_{i=0}^{M-2} f_i + M f_{M-1}\right\}$$

c) $$H_3^M(v_1, v_2) = (v_1 + 1)(v_2 + v_1)\left\{\left[v_2\sum_{i=1}^{M}\sum_{j=i}^{M} i v_1^{j-i} f_j - \sum_{i=0}^{M-1} f_i + (M + 1) f_M\right](v_2 + v_1 - 1)^2\right.$$
$$\left. + [(M + 2)(v_2 + v_1 - 1) + 1](f_M - f_{M-1})\right\}$$

d) $$H_4^M(v_1, v_2) = -(v_2 + v_1)\left[(v_1^2 - v_2^2)f_M - v_1 f_{M-1} + v_2(v_2 + v_1 - 1)\sum_{i=0}^{M-1}\sum_{j=0}^{i} v_1^{j+1} f_i\right]$$

e) $$H_5^M(v_1, v_2) = -v_2(v_2 + v_1 - 1)\left[v_2\sum_{i=0}^{M}\sum_{j=0}^{i} v_1^j f_i - f_M\sum_{i=1}^{M+1} v_1^i + f_{M-1}\sum_{i=0}^{M+1} v_1^i\right]$$

f) $$H_6^M(v_1, v_2) = (v_1 + 1)(v_2 + v_1)\left[\left(v_2\sum_{i=0}^{M}\sum_{j=0}^{i} v_1^j f_i + f_M\right)(v_2 + v_1 - 1) + f_M - f_{M-1}\right]$$

In this appendix, eqn (1a) is derived. The derivations of eqns (1b) and (1c) proceed in a very similar manner, and so are not presented in detail.

The derivation of eqn (1a) begins by noting that $(v_1 + 1)(v_2 + v_1)^2$ factors out immediately from $H_1 H_6 - H_3 H_4$, leaving:

$$-\Bigg\{(v_2 + v_1 - 1)^2\left[v_2\sum_{i=1}^{M-1}\sum_{j=i}^{M-1} i v_1^{j-i+1} f_j - v_1\sum_{i=0}^{M-1} f_i\right]$$

$$+ (v_1^2 - v_2^2)[(M + 1)(v_2 + v_1 - 1) + 1]f_M - v_1[(M + 2)(v_2 + v_1 - 1) + 1]f_{M-1}\Bigg\}$$

$$\times\left[\left(v_2\sum_{i=0}^{M}\sum_{j=0}^{i} v_1^j f_i + f_M\right)(v_2 + v_1 - 1) + f_M - f_{M-1}\right]$$

$$+\Bigg\{\left[v_2\sum_{i=1}^{M}\sum_{j=i}^{M} i v_1^{j-i} f_j - \sum_{i=0}^{M-1} f_i + (M + 1)f_M\right](v_2 + v_1 - 1)^2$$

$$+ [(M + 2)(v_2 + v_1 - 1) + 1](f_M - f_{M-1})\Bigg\}$$

$$\times\left[(v_1^2 - v_2^2)f_M - v_1 f_{M-1} + v_2(v_2 + v_1 - 1)\sum_{i=0}^{M-1}\sum_{j=0}^{i} v_1^{j+1} f_i\right] \tag{4}$$

At this point two functions are defined to simplify the presentation:

a) $$\theta_M \equiv \theta_M(v_1, v_2) = v_2\sum_{i=1}^{M}\sum_{j=i}^{M} i v_1^{j-i} f_j - \sum_{i=0}^{M} f_i \tag{5}$$

b) $$\varphi_M \equiv \varphi_M(v_1, v_2) = v_2\sum_{i=0}^{M}\sum_{j=0}^{i} v_1^j f_i + f_M$$

Substituting these functions into expression (4) and factoring out $(v_2 + v_1 - 1)$ leaves:

$$\{(M+2)v_1(v_2+v_1-1)^2(\varphi_{M-1}-f_{M-1}) - (v_1^2-v_2^2)\varphi_M + v_1(\varphi_{M-1}-f_{M-1}) + (v_1^2-v_2^2)(f_M-f_{M-1})$$

$$-(v_2+v_1-1)[v_1\theta_{M-1} + (M+1)(v_1^2-v_2^2)\varphi_M - (v_1^2-v_2^2)(\theta_M + (M+2)f_M) + 2(M+2)v_1f_{M-1} - (M+2)v_1\varphi_{M-1}]\}f_M$$

$$-(v_2+v_1-1)^2v_1[\theta_{M-1}\varphi_M - \theta_M\varphi_{M-1} + \theta_M f_{M-1}] - (v_2+v_1-1)(\theta_M-\theta_{M-1})v_1f_{M-1}$$

$$+[(M+2)(v_2+v_1-1)+1](\varphi_M-\varphi_{M-1}+f_{M-1})v_1f_{M-1} \qquad (6)$$

The following three identities, which follow directly from eqns (5) serve to complete the derivation:

$$\text{a)}\ \theta_M - \theta_{M-1} = \left(v_2\sum_{j=1}^{M} jv_1^{M-j} - 1\right) f_M \qquad (7)$$

$$\text{b)}\ \varphi_M - \varphi_{M-1} = \left(v_2\sum_{j=0}^{M} v_1^{j} + 1\right) f_M - f_{M-1}$$

$$\text{c)}\ \theta_M\varphi_{M-1} - \theta_{M-1}\varphi_M = \left[\left(v_2\sum_{j=1}^{M} jv_1^{M-j} - 1\right)\varphi_{M-1} - \left(v_2\sum_{j=0}^{M} v_1^{j} + 1\right)\theta_{M-1}\right] f_M + \theta_{M-1}f_{M-1}$$

Substituting these identities into expression (6) and factoring out $f_M$ leaves:

$$(M+2)v_1(v_2+v_1-1)^2(\varphi_{M-1}-f_{M-1}) - (v_2+v_1-1)[v_1\theta_{M-1} + (M+1)(v_1^2-v_2^2)\varphi_M]$$

$$+(v_2+v_1-1)\{[\theta_M + (M+2)f_M](v_1^2-v_2^2) - 2(M+2)v_1f_{M-1} + (M+2)v_1\varphi_{M-1}\}$$

$$-(v_1^2-v_2^2)\varphi_M + v_1(\varphi_{M-1}-f_{M-1}) + (v_1^2-v_2^2)(f_M-f_{M-1})$$

$$+(v_2+v_1-1)^2v_1\left[\left(v_2\sum_{j=1}^{M} jv_1^{M-j} - 1\right)\varphi_{M-1} - \left(v_2\sum_{j=0}^{M} v_1^{j} + 1\right)\theta_{M-1} - \left(v_2\sum_{j=1}^{M} jv_1^{M-j} - 1\right) f_{M-1}\right]$$

$$-(v_2+v_1-1)v_1\left(v_2\sum_{j=1}^{M} jv_1^{M-j} - 1\right) f_{M-1} + v_1[(M+2)(v_2+v_1-1)+1]\left(v_2\sum_{j=0}^{M} v_1^{j} + 1\right) f_{M-1} \qquad (8)$$

Substituting from eqns (5) for $\theta_M$ and $\varphi_M$ and grouping terms yields $F_M(v_1, v_2)$ as given in eqn (2) directly. This completes the derivation.

## APPENDIX E

## DERIVATION OF M = 0 NON-PREEMPTIVE SOLUTION

The mathematical interpretation of the M = 0 threshold as presented in chapter 4 results in a preemptive discipline. The system in state (0, 11) will switch to state (0, 01) at a rate of $v_1 + (1 - \beta)v_2$, implying that if server 2 terminates a service period, it will preempt the task of server 1 with probability $(1 - \beta)v_2$. In this appendix, a physical rather than mathematical extrapolation to the M = 0 case is considered and it is shown that lemma 4.2-1 holds for the physical extrapolation as well.

The state transition rate diagram for the physical extrapolation is:

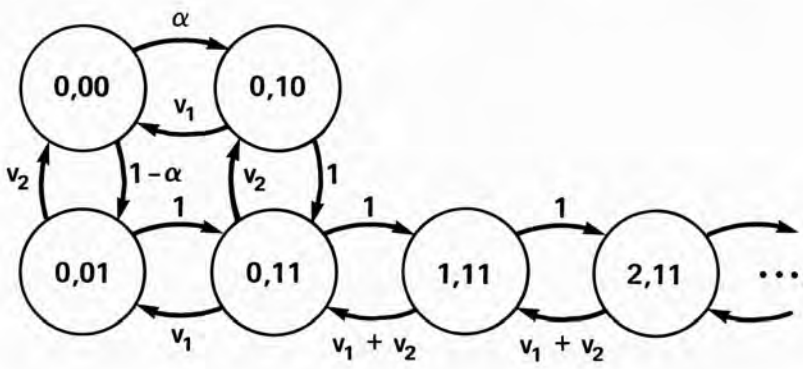



Figure E-1. State Transition Rate Diagram for Non-Preemptive M = 0 Case

The steady state equations are:

a) $p_{0,00} = v_1 p_{0,10} + v_2 p_{0,01}$ (1)

b) $(v_1 + 1)p_{0,10} = \alpha p_{0,00} + v_2 p_{0,11}$

c) $(v_2 + 1)p_{0,01} = (1 - \alpha)p_{0,00} + v_1 p_{0,11}$

d) $(v_2 + v_1 + 1)p_{0,11} = p_{0,10} + p_{0,01} + (v_2 + v_1)p_{1,11}$

e) $p_{>0,11} = \dfrac{1}{v_2 + v_1 - 1} p_{0,11}$

f) $p_{0,00} + p_{0,10} + p_{0,01} + p_{0,11} + p_{>0,11} = 1$

The solution to these equations is:

$$\text{a) } p_{0,00} = v_1 v_2 [(v_2 + v_1)^2 + v_2 + v_1 - 2]\, p_{NORM} \tag{2}$$

$$\text{b) } p_{0,10} = v_2 [\alpha(v_2 + v_1)^2 + (1 - \alpha)(v_2 + v_1) - 1]\, p_{NORM}$$

$$\text{c) } p_{0,01} = v_1 [(1 - \alpha)(v_2 + v_1)^2 + \alpha(v_2 + v_1) + 1]\, p_{NORM}$$

$$\text{d) } p_{0,11} = [\alpha v_2^2 + (v_1 + 1 - \alpha)v_2 + (1 - \alpha)v_1^2 + \alpha v_1 - 1]\, p_{NORM}$$

$$\text{e) } p_{>0,11} = [\alpha v_2 + (1 - \alpha)v_1 + 1]\, p_{NORM}$$

$$\text{f) } p_{NORM} = \{(v_1 + \alpha)v_2^3 + [2v_1^2 + (\alpha + 2)v_1 + 1]v_2^2 + (v_1 + 3 - \alpha)v_1^2 v_2 + (1 - \alpha)v_1^3 + v_1^2\}^{-1}$$

The mean number of customers in the system ($\bar{N}_0$) is:

$$\bar{N}_0 = \frac{(v_2 + v_1)^3 [\alpha(v_2 - v_1) + v_1 + 1]}{(v_2 + v_1 - 1)[(v_2 + v_1)^2(v_2 - v_1)\alpha + v_1 v_2(v_2 + v_1 + 1)^2 + (v_2 + v_1)^2 + v_1^2(v_2 + v_1) - 3v_1 v_2]} \tag{3}$$

and the partial derivative of $\bar{N}_0$ with respect to $\alpha$ is:

$$\frac{\partial \bar{N}_0}{\partial \alpha} = \frac{v_1 v_2 (v_2 - v_1)(v_2 + v_1)^3 (v_2 + v_1 + 2)}{[(v_2 + v_1)^2(v_2 - v_1)\alpha + v_1 v_2(v_2 + v_1 + 1)^2 + (v_2 + v_1)^2 + v_1^2(v_2 + v_1) - 3v_1 v_2]^2} \tag{4}$$

which is strictly negative for $v_1 > v_2$. This validates theorem 4.1-1 for the non-preemptive M = 0 case, i.e., that a probabilistic decision rule is suboptimal. The resultant deterministic decision rule yields:

$$\bar{N}_0 = \frac{(v_2 + v_1)^3 (v_1 + 1)}{(v_2 + v_1 - 1)[v_1 v_2(v_2 + v_1 + 1)^2 + (v_2 + v_1)^2 + v_1^2(v_2 + v_1) - 3v_1 v_2]} \tag{5}$$

for which we have the following corollary to lemma 4.2-1:

Lemma E-1 A threshold value of M = 1 results in a lower $\bar{N}$ than M = 0 (non-preemptive) for an ergodic system with $v_1 > v_2$.

Proof We need to show that $\bar{N}_0$ as given by eqn (5) is greater than $\bar{N}_1$ as given by eqn (4.2-2b). This is equivalent to showing:

$$(v_1 + 1)(v_2 + v_1)\left\{v_1 + v_2(v_2 + v_1) + \frac{v_2}{v_1}[(v_2 + v_1 - 1)(v_1 + 1) + 1](v_2 + v_1 + 1)\right\} >$$

$$\left[1 + \frac{v_2}{v_1}(v_2 + v_1 + 1)\right][v_1 v_2 (v_2 + v_1)^2 + v_1^3 + 3v_1^2 v_2 + 2v_1 v_2^2 + v_2^2 + v_1^2] \qquad (6)$$

This inequality can be simplified to the form:

$$v_2(v_1 - v_2)(v_2 + v_1 - 1)(v_2 + v_1)(v_2 + v_1 + 2) > 0 \qquad (7)$$

which is clearly true, thus proving the lemma.

## APPENDIX F

## DERIVATION OF LOAD-DEPENDENT TWO-SERVER SOLUTION

The load-dependent two-server system is represented by the following state transition rate diagram.

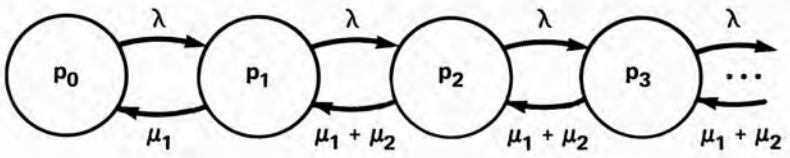



Figure F-1. State transition rate diagram for load-dependent case

The steady state equations are:

a) $\lambda p_0 = \mu_1 p_1$ (1)

b) $(\lambda + \mu_1)p_1 = \lambda p_0 + (\mu_1 + \mu_2)p_2$

c) $(\lambda + \mu_1 + \mu_2)p_i = \lambda p_{i-1} + (\mu_1 + \mu_2)p_{i+1} \quad i \geq 2$

The solution to these equations is:

a) $p_1 = \frac{\lambda}{\mu_1} p_0$ (2)

b) $p_2 = \frac{\lambda^2}{\mu_1(\mu_1 + \mu_2)} p_0$

c) $p_i = \frac{\lambda}{\mu_1 + \mu_2} p_{i-1} \quad i \geq 2$

d) $p_{>1} = \frac{\lambda}{\mu_2 + \mu_1 - \lambda} p_1$

The mean number of customers in the system during steady state operation is, therefore:

$$\bar{N} = \sum_{i=1}^{\infty} i p_i = \frac{\lambda}{\mu_1}\left[1 + 2\left(\frac{\lambda}{\mu_1 + \mu_2}\right) + 3\left(\frac{\lambda}{\mu_1 + \mu_2}\right)^2 + \ldots\right] p_0 \quad (3)$$

$$= \frac{\lambda}{\mu_1}\left[1 - \left(\frac{\lambda}{\mu_1 + \mu_2}\right)\right]^{-2} p_0 = \frac{\lambda}{\mu_1}\left(\frac{\mu_1 + \mu_2}{\mu_1 + \mu_2 - \lambda}\right)^2 p_0$$

The solution is complete with the derivation of $p_0$:

$$\sum_{i=0}^{\infty} p_i = 1 = \left[ 1 + \frac{\lambda}{\mu_1} + \frac{\lambda^2}{\mu_1(\mu_2 + \mu_1 - \lambda)} \right] p_0$$

$$= \left[ \frac{\mu_1^2 + \mu_1\mu_2 + \lambda\mu_2}{\mu_1(\mu_2 + \mu_1 - \lambda)} \right] p_0 \qquad (4)$$

And we have:

$$p_0 = \frac{\mu_1(\mu_2 + \mu_1 - \lambda)}{\mu_1^2 + \mu_1\mu_2 + \lambda\mu_2} \qquad (5)$$

References

[AGNEC76] Agnew, C. E. On quadratic adaptive routing algorithms. Comm. ACM 19, 1 (Jan. 1976), 18–22.

[ALLEA75] Allen, Arnold O. Elements of queueing theory for system design. IBM Systems Journal 14, 2 (1975), 161–187.

[ALLEA78] Allen, Arnold O., Probability, Statistics, and Queueing Theory: With Computer Science Applications, Academic Press, New York (1978).

[ALLEA80] Allen, Arnold O. Queueing models of computer systems. Computer 13, 4 (April 1980), 13–24.

[BELLC75] Bell, Colin E. Turning off a server with customers present: Is this any way to run an M/M/c queue with removable servers? Operations Research 23, 3 (May–June 1975), 571–574.

[BELLC80] Bell, Colin E. Optimal operation of an M/M/2 queue with removable servers. Operations Research 28, 5 (Sept.–Oct. 1980), 1189–1204.

[BHATU68] Beckman, M. and Kunzi, H. P., eds. Lecture Notes in Operations Research and Mathematical Economics, Springer–Verlag, Berlin (1968), Vol. 2: A Study of the Queueing Systems M/G/1 and GI/M/1, by U. Narayan Bhat.

[BOXMO79] Boxma, O. J., Cohen, J. W., and Huffels, N. Approximations of the mean waiting time in an M/G/s queueing system. Operations Research 27, 6 (Nov.–Dec. 1979), 1115–1127.

[BOYSJ75] Boyse, John W. and Warn, David R. A straightforward model for computer performance prediction. Computing Surveys 7, 2 (June 1975), 73–93.

[BUZEJ73] Buzen, Jeffrey P. Computational algorithms for closed queueing networks with exponential servers. Comm. ACM 16, 9 (Sept. 1973), 527–531.

[CHANK72] Chandy, K. M., Keller, T. W., and Browne, J. C. Design automation and queueing networks: An interactive system for the evaluation of computer queueing models. Proc. of the Ninth Design Automation Workshop, (June 1972), 357–367.

[CHANK75] Chandy, K. M., Sauer, C. H., and Browne, J. C. An overview of modeling techniques for parallel processing systems. Digest of Papers: Compcon Spring 75, Tenth IEEE Computer Society International Conference, 213–218.

[CHENP73] Chen, Peter P. S. Optimal file allocation in multi-level storage systems. Proc. National Computer Conference, 1973, 277–282.

[CHENP75] Chen, Peter P. S. Queueing network model of interactive computing systems. Proc. of the IEEE 63, 6 (June 1975), 954–957.

[CHOWY79] Chow, Yuan-Chieh and Kohler, Walter H. Models for dynamic load balancing in a heterogeneous multiple processor system. IEEE Trans. on Computers C-28, 5 (May 1979), 354–361.

[COFFE73] Coffman, Edward G., Jr. and Denning, Peter J., Operating Systems Theory, Prentice-Hall, Inc., Englewood Cliffs, New Jersey (1973).

[COHEJ76] Cohen, J. W. On the optimal switching level for an M/G/1 queueing system. Stochastic Processes and Their Applications 4 (1976), 297–316.

[CRABT77] Crabill, Thomas B., Gross, Donald, and Magazine, Michael J. A classified bibliography of research on optimal design and control of queues. Operations Research 25, 2 (March-April 1977), 219–232.

[DENNP76] Denning, Peter J. and Kahn, Kevin C. An L = S criterion for optimal multiprogramming. Proc. of the International Symposium on Computer Performance Modeling, Measurement and Evaluation, Harvard University, Cambridge, MA, (March 29–31, 1976), 219–229.

[DERMC70] Derman, Cyrus, Finite State Markovian Decision Processes, Academic Press, New York (1970).

[FAYOG76] Fayolle, G., and Robin, M. Optimal queueing policies in multiple-processor computers. In Modelling and Performance Evaluation of Computer Systems, pp. 103–117. Edited by E. Gelenbe and H. Beilner. North-Holland Publishing Company (1976).

[GAVED67] Gaver, D. P., Jr. Probability models for multiprogramming computer systems. JACM 14, 3 (July 1967), 423–438.

[HOWAR71] Howard, Ronald A., Dynamic Probabilistic Systems, Volume I: Markov Models, John Wiley and Sons, Inc., New York (1971).

[HEYMD80] Heyman, Daniel P. Comments on a queueing inequality. Management Science 26, 9 (Sept. 1980), 956–959.

[KLEIL70] Kleinrock, Leonard. Survey of analytical methods in queueing networks. In Computer Networks, pp. 185–205. Edited by Randall Rustin. Prentice-Hall, N.J. (1970).

[KLEIL75] Kleinrock, Leonard, Queueing Systems, Vol. 1: Theory, John Wiley and Sons, Inc., New York (1975).

[KLEIL76] Kleinrock, Leonard, Queueing Systems, Vol. 2: Computer Applications, John Wiley and Sons, Inc., New York (1976).

[KOBAH77] Kobayashi, Hisashi and Konheim, Alan G. Queueing models for computer communications system analysis. IEEE Trans. on Communications COM-25, 1 (Jan. 1977), 2-29.

[LEROJ76] Leroudier, Jacques and Potier, Dominique. Principles of optimality for multiprogramming. Proc. of the International Symposium on Computer Performance Modeling, Measurement and Evaluation, Harvard University, Cambridge, MA, (March 29-31, 1976), 211-218.

[LEVYY76] Levy, Yonatan and Yechiali, Uri. An M/M/s queue with servers' vacations. INFOR 14, 2 (June 1976), 153-163.

[LITTJ61] Little, J. D. C. A proof for the queueing formula: $L = \lambda W$. Operations Research 9, 3 (May-June 1961), 383-387.

[MAGAM71] Magazine, M. J. Optimal control of multi-channel service systems. Naval Research Logistics Quarterly 18 (1971), 177-183.

[MARTJ78] Martin, James, Communications Satellite Systems, Prentice-Hall, Inc., Englewood Cliffs, N. J. (1978).

[REISM76] Reiser, M. and Konheim, A. G. Finite capacity queueing systems with applications in computer modeling. Report RC5827, IBM Thomas J. Watson Research Center (1976).

[RICAG80] Ricart, Glenn. Efficient synchronization algorithms for distributed systems. Ph.D. dissertation, Tech. Rpt. TR-902, Dept. of Computer Science, Univ. of Maryland (May 1980).

[ROSEM75] Rosenshine, Matthew. Queueing theory: The state of the art. AIIE Transactions 7, 3 (Sept. 1975), 257-267.

[SAKAH77] Sakasegawa, Hirotaka. An approximation formula $L_q \simeq \alpha \cdot \rho^{\beta}/(1 - \rho)$. Annals of the Institute of Statistical Mathematics 29, 1 (1977), Part A, 67–75.

[SASTK75] Sastry, K. V. and Kain, R. Y. On the performance of certain multiprocessor computer organizations. IEEE Trans. on Computers C–24, 11 (Nov. 1975), 1066–1074.

[SERFR79] Serfozo, Richard F. An equivalence between continuous and discrete time Markov decision processes. Operations Research 27, 3 (May–June 1979), 616–620.

[SHANJ79] Shanthikumar, J. G. On a single–server queue with state–dependent service. Naval Research Logistics Quarterly 26, 2 (June 1979), 305–309.

[SOBEM80] Sobel, Matthew J. Simple inequalities for multiserver queues. Management Science 26, 9 (Sept. 1980), 951–956.

[STIDS73] Stidham, S., Jr. and Prabhu, N. U. Optimal control of queueing systems. In Mathematical Methods in Queueing Theory, pp. 263–294. Edited by A. B. Clarke. Springer–Verlag, New York (1973).

[STRAJ74] Strauss, J. C. An analytic model of the HASP execution task monitor. Comm. ACM 17, 12 (Dec. 1974), 679–685.

[TOWSD80] Towsley, Don. Queueing network models with state–dependent routing. JACM 27, 2 (April 1980), 323–337.

[TRIVK80] Trivedi, Kishor S., Wagner, Robert A., and Sigmon, Timothy M. Optimal selection of CPU speed, device capacities, and file assignments. JACM 27, 3 (July 1980), 457–473.

[WEBER80] Weber, Richard R. On the marginal benefit of adding servers to G/GI/m queues. Management Science 36, 9 (Sept. 1980), 946–951.

[WINSW77] Winston, Wayne L. Assignment of customers to servers in a heterogeneous queueing system with switching. Operations Research 25, 3 (May–June 1977), 469–483.

[YADIM67] Yadin, M. and Naor, P. On queueing systems with variable service capacities. Naval Research Logistics Quarterly 14, 1 (March 1967), 43–53.

## BIBLIOGRAPHIC DATA SHEET

| 1. Report No. | 2. Government Accession No. | 3. Recipient's Catalog No. |
|---|---|---|
| TM 82119 | | |

| 4. Title and Subtitle | 5. Report Date |
|---|---|
| Control of Multiple Exponential Servers with Application to Computer Systems | April 1981 |
| | 6. Performing Organization Code |

| 7. Author(s) | 8. Performing Organization Report No. |
|---|---|
| Ronald L. Larsen | |
| | 10. Work Unit No. |

| 9. Performing Organization Name and Address | 11. Contract or Grant No. |
|---|---|
| Ground Systems Management Office<br>Goddard Space Flight Center<br>Greenbelt, Maryland 20771 | |
| | 13. Type of Report and Period Covered |
| 12. Sponsoring Agency Name and Address | Technical Memorandum |
| | 14. Sponsoring Agency Code |

15. Supplementary Notes

Ph.D. Dissertation submitted to the Department of Computer Science, University of Maryland, College Park, MD 20742. Copies available through University Microfilms International.

16. Abstract

A class of dynamic control policies is defined for scheduling customers from a Poisson source on a set of exponential servers with dissimilar service rates. A fastest-to-slowest ordering is imposed on the servers, and they are invoked in response to instantaneous system loading as measured by the length of the queue of waiting customers. Markov chain analysis is employed to analyze the performance of the control policies and to develop optimality criteria. It is shown for the two-server case, and believed to be true in general, that probabilistic control policies are suboptimal to minimize the mean number of customers in the system, and that the optimal policy is in a restricted class of policies referred to as "threshold queueing" policies. In a threshold queueing policy, specific queue lengths are identified as "thresholds" beyond which an additional (fastest idle) server is invoked. A new server remains busy until it completes service on a customer and the queue length is less than its invocation threshold.

| 17. Key Words (Selected by Author(s)) | 18. Distribution Statement |
|---|---|
| Queueing Theory<br>Resource Allocation<br>Scheduling<br>Multiple Servers | |

| 19. Security Classif. (of this report) | 20. Security Classif. (of this page) | 21. No. of Pages | 22. Price* |
|---|---|---|---|
| Unclassified | Unclassified | | |

*For sale by the National Technical Information Service, Springfield, Virginia 22161.

16. Abstract (continued)

Optimality conditions are derived, and an approximation to the optimum policy is analyzed. For most operational applications, a very simple approximation of the optimal threshold value suffices.

Extensions to the basic policy involving different objectives are considered, including a treatment of the n-server case, the finite queue situation, and inverse ordering (slowest-to-fastest) of servers. Several applications of threshold queueing policies in computer and communications systems are presented. It is concluded that threshold queueing policies with easily approximated thresholds provide near-optimal control of multiple exponential servers wtih dissimilar service rates, and that these policies can be readily applied to improve the performance of contemporary computer and communications systems.